THIS WEEK

EDITORIALS

Don't rush to rehabilitate Hwang

Nature's profile of a former fraudster's attempts to regain respectability should not be taken as an endorsement of the researcher's claims.

An article published on *Nature*'s website last week has created quite a buzz in South Korea. It details efforts by former Seoul National University cloning specialist Woo Suk Hwang to rehabilitate his scientific career after he was found in 2006 to have been involved in fraud. Some in South Korea are taking the article as a sign that Hwang is now producing great science and is once again lauded by the scientific community. Stock prices of companies with connections to Hwang's work have apparently jumped. It is as if many of the people talking and writing about the article have not read it. They and others can do so now if they wish: it appears as a News Feature on page 468.

As readers will see, the article is not a show of support for Hwang's research. Nor is it an attack. It is the story of a rare event: a scientist attempting with some success to dig himself out from the depths of ignominy. It is a journalistic exercise, not a scientific endorsement. And it was commissioned to mark the ten-year anniversary of the first paper — now retracted — in which Hwang claimed to have created cloned human embryonic stem-cell lines.

The article highlights notes of caution for those who would rush to rehabilitate this disgraced researcher. Most worryingly, Hwang is pushing — with some success — to get recognition that his cells are indeed the world's first cloned human stem-cell line. That is not supported either by independent scientific evidence produced since he published his now-retracted paper, or by evidence from his own laboratory, which fabricated data after tests showed that the cell line was not cloned. Hwang has taken the unscientific path of getting patent offices and court rooms, rather than his expert peers, to judge his scientific claims.

Hwang's position panders to the views of many of his diehard supporters, who treat the matter as if a great scientist's great discovery had been somehow unfairly taken away; as if Hwang lost his reputation on a technicality. Indeed, the whistle-blower who endured persecution to set the record straight about Hwang's research has been portrayed online as a traitor who embarrassed the country, hampered a distinguished scientist and set back the progress of South Korea's biotechnology.

Nothing could be further from the truth. The evidence suggests that Hwang was not a great scientist. His claims to have done cloning work on cows in the late 1990s were backed up with photographs and promoted through political connections rather than scientific publications. What was the contribution to scientific knowledge of his human-cloning work? In May 2013, cell biologist Shoukhrat Mitalipov published results showing that he had finally achieved the human-cloning breakthrough that Hwang had claimed in 2004. Mitalipov told *Nature*: "I don't have much to say about Hwang; his studies in human somatic-cell nuclear transfer were not informative and did not affect me at all." Eggs were given in vain to Hwang's lab by around 120 donors. The potential of Hwang's claimed work was over-hyped even before the work was exposed as fraudulent, especially considering that superior technologies — such as stem cells made from reprogrammed adult cells — were already in the offing.

The whistle-blower did not cause South Korea to lose anything. There was nothing to lose. What he did was cut short attempts to trumpet overblown and dishonest research. He helped to nip misguided efforts in the bud so that South Korean science could move on.

And it has. Undeterred by the Hwang scandal, the government has invested generously in stem cells and other scientific fields. The country's current work might not be that earthshaking, but great breakthroughs often come when one is neither expecting nor promising much.

"The whistle-blower helped to nip misguided efforts in the bud so that South Korean science could move on."

If Hwang wants to rebuild his scientific reputation, which he seems intent on doing, and which his scientific colleagues seem willing to accept — some grudgingly — a good start would be jettisoning his patent claims and other legal efforts to be recognized as having created the first cloned human stem-cell line. People are asking, can we trust him? Part of the answer lies in how he resolves this issue. If he wants to start again, he should look there. ■

A return to order

Members of the US Congress have taken a much-needed step to restore credibility.

There is big excitement on the US political scene this week with the news that Congress has finally passed a budget to fund the government for the remainder of fiscal year 2014 (see page 461). The good news for US scientists is that support for their work remains strong: most research-funding agencies (with the notable exception of the National Institutes of Health) have seen a partial restoration of funding after the across-the-board cuts mandated last year under the sequester.

But the better news for everyone is the existence of the settlement itself: it marks the first return in years to anything resembling a normal budget process. Given the poisonous partisanship that has dominated US politics in recent years, the simple act of funding the government — achieving what any other country would consider routine — has required gruelling negotiations and rare political courage. Better still, the success of those efforts offers at least some hope that they will be repeated in future years — that the stranglehold of the uncompromising, anti-government, largely Republican minority known as the Tea Party has at last been broken.

Credit for the budget's success goes in the first instance to

Representative Paul Ryan (Republican, Wisconsin) and Senator Patty Murray (Democrat, Washington), chairs of the House and Senate budget committees, respectively. In the aftermath of last autumn's government shutdown, Ryan and Murray negotiated an overall budget figure of US$1.1 trillion that eases some of the sequester cuts beloved of Republicans while excluding unemployment benefits favoured by Democrats. In the process, they faced down claims of betrayal from both sides.

Arguably, even more credit is due to Representative Hal Rogers (Republican, Kentucky), who chairs the House appropriations committee, and Senator Barbara Mikulski (Democrat, Maryland), chair of the Senate equivalent. They had the unenviable job of allocating the overall budget among specific departments and programmes. These two politicians disagree on almost every issue. But they had the sense and judgement to agree on this: even leaving aside the disaster of the shutdown and the mindlessness of the sequester, Congress cannot keep on funding the government year to year with 'continuing resolutions' that avoid making choices, and instead keep programmes going on a yearly basis just as they were. The result is waste, turmoil and missed opportunities in the agencies, in which demoralized officials are forced to defer long-planned initiatives, hoard the money they do have and spend their days endlessly planning and replanning.

Once Ryan and Murray's overall budget number was in hand, Rogers, Mikulski and their staff worked almost non-stop to agree on allocations. They both had to deal with members of their own parties who wanted to attach amendments promoting this or that pet cause, and plenty of those measures did make it into the final bill. But they managed to fend off the worst of the 'poison pill' amendments that were designed to force the opposite party to vote against the final package — including one that would have blocked the Environmental Protection Agency from regulating greenhouse-gas emissions to fight climate change.

Finally, credit is due to the rank-and-file members of Congress who passed the budget bill by overwhelming bipartisan majorities — despite threats from staunchly partisan political groups to use those votes against members in the upcoming November elections, when every Representative and one-third of Senators will face the voters.

The problem now is that the current spending agreement runs only until 1 October, the start of fiscal year 2015. If no new overall budget is agreed, the dreaded sequester will return, and with it the automatic, widespread budget reductions totalling roughly $100 billion every year until 2023. Rogers and Mikulski have vowed to pursue a new agreement as part of their wider intention to continue Congress' return to normal procedure. But in an election year, it is not clear whether they will have enough time. To allow everyone ample opportunity to campaign, the House is scheduled to be in session for only 113 days this year, and the Senate for just less than 200. Nor is it clear how much courage anyone will maintain once the negative campaign adverts start flying.

"It may come down to how well Congress learned the lessons of the shutdown."

Ultimately, it may come down to how well individual members of Congress learned the lessons of the shutdown, which sent public-approval ratings for their institution spiralling into the single digits late last year. If memories are short, and members sink back into a miasma of mistrust and gridlock, then sequestration looms.

Instead, if they can return to behaving like rational adults, then there is hope. Perhaps Congress can start making the kind of investments in research, education and infrastructure, such as broadband and smart grids, that both parties say are needed — and that foster the kind of economic growth that both parties say they want. ■

A question of time

Timekeeping is boosted by the advent of an optical clock based on strontium atoms.

When the history of the twenty-first century comes to be written, one of the most puzzling questions asked will be why, well into the information age, millions of people still paid to dial a number on their phone to find out the time. Almost 80 years after its formation, the UK speaking clock, the world's original telephone time service, remains an essential part of British life. This is despite the near ubiquity of time displays — not least on the mobile phones that people discard to call 123 from a fixed line.

For some people, at some times, accuracy matters. Peaks in the use of the speaking clock come, for instance, on New Year's Eve, or when the clocks are put forward and back by an hour to mark, respectively, the start and end of British Summer Time.

There is another way, at least in Britain. BBC Radio regularly broadcasts the same time signal used to set the speaking clock — affectionately known as the pips. Indeed, it has become as much a feature of some shows as the content planned around it. Time is more than a British institution; it is woven into the cultural fabric of everyday life.

The pips are drawn from an atomic clock held at the National Physical Laboratory (NPL) in Teddington, near London. One of the most accurate in the world, the NPL clock is tuned to the regular bursts of light emitted by caesium atoms when they are excited by microwaves. The clock would lose roughly one second every 138 million years — a sufficient degree of accuracy for a bleary-eyed hour-late commuter who forgot to set their clock the night before, but not accurate enough for some.

In a paper published on *Nature*'s website this week, time lords in the United States describe the latest advance in chronometry, and one that is as superior to the atomic pips as those pips were to the mechanical devices they replaced (B. J. Bloom *et al. Nature* http://dx.doi.org/10.1038/nature12941; 2014). The researchers have built a timepiece based not on caesium but on strontium. More importantly, it uses much higher, optical frequencies. This gives such devices, called optical clocks, greater accuracy than those that rely on microwaves. The new optical clock, for example, would not lose one second even if it were to run for 5 billion years.

It is also extremely stable — another key measure of timekeeping. (Accuracy defines how closely a clock's output matches the desired time signal, whereas stability is a measure of how steady that output is. A clock that loses precisely one second each day is inaccurate but stable, for example.)

The unveiling of the super-accurate strontium optical clock comes just a few months after a related group revealed a device based on ytterbium. Other laboratories across the world have their own designs. Inevitably, the increased precision and reliability of optical clocks are fuelling debate about whether they could be used to set the ultimate time, and redefine the second. (There are no official plans to do so, but plans are afoot to redefine other SI units.) These are heady times for metrology: a World View on page 455 describes attempts to measure another fundamental constant: Big *G*.

Nature has a particular stake in the race to develop new atomic clocks. Back in January 2003, we published a News Feature that surveyed the scene and tried to predict what would happen (D. Adam *Nature* **421,** 207–208; 2003). Within a decade, the piece suggested, optical clocks could rise to prominence and raise fresh debate about the definition of the second. A ten-year event horizon is a staple of scientific journalism, and most promised breakthroughs fail to materialize on deadline. The latest development in atomic timekeeping, by contrast, has arrived bang on time. Well, almost. ■

➲ NATURE.COM
To comment online, click on Editorials at: go.nature.com/xhunqv

Don’t stop the quest to measure Big *G*

Pinning down the coupling constant in Newton’s law of gravity is challenging, but with ultra-stable labs it can be done, says **Terry Quinn**.

Plans are well under way to redefine the International System of Units (SI) by basing it on seven physical constants — the kilogram, for instance, is to be linked to the numerical value of Planck’s constant, rather than the current definition of the mass of a platinum–iridium cylinder.

The seven constants help to explain and predict the motion and actions of the Universe. But one that is not included is what we in the business call Big *G*: the coupling constant in Newton’s law of gravity.

There is a problem with Big *G* (so called to distinguish it from little *g*, the acceleration due to gravity at Earth’s surface). Current measurements of it are, frankly, all over the place. Seven separate experiments in the past decade or so have given results that have a spread of about 0.05%. For a fundamental constant of physics, that is extraordinarily imprecise.

This uncertainty has little impact on day-to-day life, and Einstein’s general theory of relativity has long replaced Newton’s law of gravity as the way for scientists to view physics at the largest scales. Yet Newton’s law still predicts with adequate precision the movements of the planets and their moons, artificial satellites and space probes. (We don’t need to isolate the value of *G* to calculate these movements, because the equations depend on a combination of their masses and *G*.)

THERE IS AN ALMOST IRRESISTIBLE PRESSURE **TO STOP** WHEN THE RESULT IS ABOUT WHAT ONE EXPECTS IT **TO BE.**

For a scientist — and a former director of the International Bureau of Weights and Measures (BIPM) in Paris such as myself — the imprecision in *G* is irritating. Moreover, there is a solid scientific case for sorting it out. The search for a theory of quantum gravity that is consistent with quantum electrodynamics is perhaps the most active field of theoretical physics. One day, we may have to test such theories by comparing the values of *G* that they predict with the real thing — so we need an accurate experimental value.

The problem for the physicist on Earth who tries to measure *G* is that, although the strength of gravity is huge on an astronomical scale, it is extremely small in a laboratory. The force of gravity holds the planets in their orbits around the Sun and the billions of stars in the arms of the galaxies, yet this is the same gravitational force that, between a pair of 1-kilogram copper balls that are just touching, is about 10^{-8} times the weight of each.

To pick up this tiny signal, the laboratory itself has to be mechanically stable, with a low level of ground vibration and tilt, and with the temperature of its apparatus stabilized to a few thousandths of a degree Celsius.

Assuming that there is no hidden physics that can explain why the value of *G* measured in different places would be different (unlikely), why is there such a spread of results? The problem lies in systematic error — the spectre that haunts every absolute determination of a fundamental constant. No matter how much one tries to take into account every possibility for error in a measurement, it is in principle impossible to demonstrate its absence. The only way to give confidence is to measure the same constant using a number of different methods. This is true in the measurement not only of a fundamental constant of nature, but of anything else.

At the BIPM, we devised an experiment to measure *G* with two almost-independent methods in the same apparatus. Our results were at the high end of the range. They did not have the smallest level of uncertainty of any *G* experiments, but they are the only ones that have been repeated and it is the only *G* experiment in which more than one method has been used (we published the original in 2001 and then the follow-up last year).

There is another, more subtle, problem, which is related to the experimenter’s behaviour. In measuring *G*, or any other constant, one starts out with a pretty good idea of its value. As the data come in, it starts to become clear roughly where the final result will lie. More data refine the value and correct both large and small errors. At what point does the experimenter stop searching for errors? There is an almost irresistible pressure to stop when the result is about what one expects it to be. Concealing the results from the experimenter solves this problem but sets up another — crude errors are missed and thus waste valuable time and effort.

So it is difficult to find the true value of *G*, but I believe that it can be done. New efforts and new measurements are needed. Both the US National Institute of Standards and Technology in Gaithersburg, Maryland, and the UK National Physical Laboratory in Teddington have or are building ultra-stable metrology laboratories that offer the best chance yet to pin down Big *G*. I have already suggested to them that they carry out *G* experiments — and, of course, I would like those to be improved versions of the BIPM experiment.

Is it worth it? The answer must be yes. I will make the case again at a meeting of physicists and metrology experts at the Royal Society in London next month, organized by me, Clive Speake of the University of Birmingham, UK, and Luo Jun of Huazhong University of Science & Technology in Wuhan, China. The title of the meeting is ‘The Newtonian constant of gravitation, a constant too difficult to measure?’. The answer to that is surely no. ■

NATURE.COM
Discuss this article online at:
go.nature.com/Gwheyh

Terry Quinn *is emeritus director of the BIPM near Paris.*
e-mail: tjqfrs@gmail.com

RESEARCH HIGHLIGHTS *Selections from the scientific literature*

VISION

Dopamine loss hurts diabetic eye

A decrease in the amount of dopamine in the retina could explain why people with diabetes often have visual problems or even go blind.

Reduced levels of this brain-signalling molecule have been seen in diabetes before, so Machelle Pardue at Emory University in Atlanta, Georgia, and her colleagues gave a dopamine precursor called L-DOPA to rat and mouse models of type 1 diabetes. They found that the molecule delayed the onset and slowed down the progression of early visual dysfunction, and improved the responses of the retina's light-sensing cells.

Treating dopamine deficiency could be a way to combat vision loss associated with type 1 diabetes, the authors say.

J. Neurosci. 34, **726–736 (2014)**

JORGE GONZALEZ VARGAS PHOTOGRAPHY/GETTY

OCEANOGRAPHY

Climate change spawns bigger waves

Taller ocean waves could slam coastal regions in the tropics and in parts of the Southern Hemisphere this century, thanks to faster surface winds.

Xiaolan Wang and her colleagues at Environment Canada in Toronto developed statistical models that use sea-level pressure data from multiple global climate model simulations to predict changes in the height of ocean waves. The authors found that the frequency of extremely high waves that now occur roughly once a decade could double or triple by the end of this century in some coastal regions, including Chile (pictured) and Mexico's Baja peninsula. Surface wind speeds are affected by changing air temperature and sea-level pressure.

Rising sea levels could worsen the impacts of bigger waves, such as coastal flooding and erosion, the authors say.

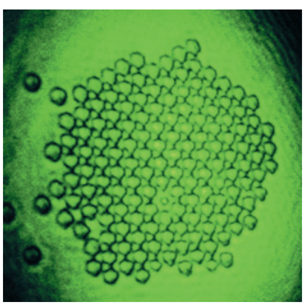
Geophys. Res. Lett. **http://doi.org/q2c (2014)**

OPTICS

Laser power makes a mirror

Physicists have created a mirror by using a laser to herd tiny particles into a continuous reflective surface.

Optical forces from laser beams have already been used to manipulate single particles. Now, Tomasz Grzegorczyk at BAE Systems in Burlington, Massachusetts, and his colleagues have used a green laser to organize about 150 polystyrene spheres suspended in water. The three-micrometre-wide beads interacted with each other to form a crystal-like membrane configuration (**pictured**), and a camera was used to capture an image reflected off the membrane surface.



The method gives the membrane self-healing properties, and one day could be used to build ultralight mirrors with a large surface area for space telescopes, the authors suggest.

Phys. Rev. Lett. 112, **023902 (2014)**

MICROBIOLOGY

How antibiotics boost infection

Antibiotics alter the bacterial community in the mouse gut in ways that might make the animal more susceptible to infections from the dangerous, diarrhoea-causing bacterium *Clostridium difficile*.

Vincent Young and his team at the University of Michigan in Ann Arbor analysed the molecules produced by gut microbes and found that antibiotics shifted the levels of carbohydrates and other metabolites.

Compounds that became more abundant with antibiotic treatment such as the sugar alcohols mannitol and sorbitol boosted the growth of *C. difficile* cells in culture. A bile acid that also increased in treated mice triggered spores of the bacterium to germinate. Moreover, the

T. M. GRZEGORCZYK (BAE SYSTEMS/MIT)/J. ROHNER/J.-M. FOURNIER (EPFL)

intestinal contents from mice given antibiotics promoted the growth of *C. difficile*, whereas those from untreated mice did not.

The results could explain why people taking antibiotics have a high risk of *C. difficile* infection.

Nature Commun. **5, 3114 (2014)**

NEUROSCIENCE

Drugs help to dull bad memories

A drug can improve the effectiveness of a behavioural treatment for fearful memories, at least in mice.

Long-term memories of traumatic events, which can result in anxiety disorders, are difficult to treat, in part because they leave epigenetic, or chemical, marks in the genome. Li-Huei Tsai at the Massachusetts Institute of Technology in Cambridge and her colleagues tested an HDAC inhibitor, a drug that clears epigenetic markers, on mice that were conditioned to freeze in fear when they heard a loud sound. Conditioned mice given the drug, and then exposed to the sound in a safe environment, froze much less frequently than mice that did not receive the drug. The inhibitor made it easier to replace the bad memory with a less fearful one by changing the expression of the genes involved in rewiring the brain, the authors say.

Cell **156, 261–276 (2014)**

ENTOMOLOGY

Parasite drives host to nectar

Mosquitoes carrying a malaria-causing parasite develop an increased desire for sugar.

Baldwyn Torto of the International Centre of Insect Physiology and Ecology in Nairobi and his colleagues monitored the attraction of *Anopheles gambiae* mosquitoes (**pictured**) to plant odours and the investigative behaviour of the insects around nectar sources. In laboratory experiments, the authors showed that insects infected with *Plasmodium falciparum* parasites were more attracted to plant odours and demonstrated increased pre-feeding probing activity compared with uninfected individuals.

Plant odours could be used to trap parasite-infected mosquitoes, the authors suggest.

Curr. Biol. **http://doi.org/qww (2014)**

KALLISTA IMAGES/GETTY

CLIMATE CHANGE

Strong storms shift landwards

Cyclone activity has shifted towards the coasts in east Asia in recent decades, resulting in storms of greater intensity making landfall over eastern China, Korea and Japan.

Chang-Hoi Ho of Seoul National University and his colleagues analysed east Asian storm data from 1977 to 2010. The frequency of intense storms that hit northerly areas has increased, but the intensity of cyclones making landfall farther south — from Vietnam to Taiwan — has not measurably changed.

The researchers suggest that changing atmospheric-circulation patterns resulting from a gradual warming of the western Pacific Ocean have shifted the areas where cyclones develop, moving them to the north and west.

Environ. Res. Lett. **9, 014008 (2014)**

IMMUNOLOGY

Microbes control immune cells

Beneficial gut bacteria secrete compounds that rein in a group of immune cells that are involved in inflammatory disorders.

Microbes in the gut help to keep immune responses in check, but how they do this has not been clear. To find out, Richard Blumberg and Dennis Kasper of Harvard Medical School in Boston, Massachusetts, and their team studied a helpful intestinal bacterium, *Bacteroides fragilis*.

Mice colonized with *B. fragilis* had fewer 'natural killer T cells' than did mice without the bacterium. But that effect was reduced in mice harbouring *B. fragilis* that lacked a gene responsible for making fatty compounds called sphingolipids.

Treating these mice with a purified *B. fragilis* sphingolipid restored normal natural killer T-cell inhibition and protected the mice from chemically induced colitis. Animals exposed to the microbe early in life were more protected than those exposed later.

Cell **156, 123–133 (2014)**

CHEMISTRY

Molecules built in a bubble

Chemical synthesis occurs more readily if the reaction takes place inside micrometre-sized compartments.

In theory, it is difficult to merge two molecules into one because of the decrease in entropy as the reaction proceeds. To overcome this hurdle, Andrew Griffiths at Strasbourg University in France and his colleagues studied chemical reactions occurring inside tiny water droplets suspended in oil.

They found that a fluorescent molecule built from two reagents formed more quickly in smaller droplets of water. A mathematical model indicated that molecules landing on a droplet's internal surface are more likely to merge with each other because the surface limits the available space and constrains the reactants' freedom of movement.

The results suggest that compartments, such as aerosol droplets or the pores in hydrothermal vents, could have assisted the organic reactions that are thought to have led to the origin of life.

Phys. Rev. Lett. **112, 028310 (2014)**

COMMUNITY CHOICE

The most viewed papers in science

APPLIED PHYSICS

Device harvests power from the air

HIGHLY READ on scitation.aip.org in December

Researchers have built a metamaterial device that captures microwaves and turns them into electrical power.

Metamaterials are made up of structures that are smaller than a given wavelength of electromagnetic radiation. When arranged in arrays, these structures can tune waves of that radiation in novel ways. Allen Hawkes and his colleagues at Duke University in Durham, North Carolina, used an array of five metamaterial cells, made from split copper circuits, to harvest microwave energy.

The microwaves cause an oscillating current in the material, and the copper circuits convert part of that current into usable power. When high-power microwaves were applied, the array produced enough direct current to charge a mobile phone.

Such a material could one day be built into devices and generate power by picking up energy from a mobile phone or Wi-Fi signals, the authors say.

Appl. Phys. Lett. **103, 163901 (2013)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

ERICH SCHLEGEL/CORBIS

Extinction risk for sharks and rays

One-quarter of all species of shark and ray are threatened by overfishing, according to the first global analysis of the animals' conservation status. The figures come from the Shark Specialist Group at the International Union for Conservation of Nature (IUCN), which analysed 1,041 species of the chondrichthyan class (sharks, rays and 'ghost sharks' called chimaeras). Of these, 25 are critically endangered, 43 endangered, 113 vulnerable and 132 'near threatened', according to IUCN criteria. Extinction risk among these animals is "substantially higher than for most other vertebrates", the group reports (N. K. Dulvy *et al. eLife* **3,** e00590; 2014). See go.nature.com/qi2tvi for more.

RESEARCH

Antarctic blast

An explosion at Argentina's Esperanza station in Antarctica killed one person on 14 January, according to news reports. The blast occurred during the handling of flammable materials. Esperanza, on the northern tip of the Antarctic Peninsula, is used for biology, geology and seismology research.

POLICY

US budget bill

On 17 January, US President Barack Obama signed a US$1.1-trillion spending bill that will fund the government until 30 September. The budget restores most science agencies to roughly 2012 funding levels — as they stood before last year's across-the-board government spending cuts, known as sequestration. But funding for the National Institutes of Health still fell $1.25 billion short of what Obama had requested for the 2014 fiscal year. See page 461 for more.

Energy trends

Global investments in renewable energy dropped for the second consecutive year last year, with a 12% dip from 2012, according to figures released on 15 January by international research company Bloomberg New Energy Finance. The falling costs of solar installations helped to drive the downturn, as did uncertain government support for renewable power in Europe and the United States. Europe's investments plummeted by US$40 billion, or 41%, from 2012. By contrast, Japan saw a 55% surge, as closures of nuclear power plants made room for the solar industry.

Bisphenol A limits

The European Food Safety Authority (EFSA) has recommended slashing daily intake limits for the chemical bisphenol A (BPA), used to manufacture plastics such as food containers. In a draft assessment released on 17 January, the agency said that BPA poses a low public-health risk because typical exposure levels are low. But uncertainty about the chemical's potential toxicity, including effects on the reproductive and nervous systems, warrants temporarily cutting daily limits from 50 to 5 micrograms per kilogram of body weight, the EFSA said.

Extreme weather

The United States experienced warmer and wetter weather than average in 2013, the National Oceanic and Atmospheric Administration reported in an annual summary on 15 January. Across the 48 contiguous states, temperatures averaged 11.3 °C last year, 0.2 °C above the twentieth-century average. Ten states reported one of their ten wettest years on record. By contrast, California recorded its driest year, with 27.6% of the state in severe drought by the end of 2013.

BUSINESS

Bargain genome

Biotechnology company Illumina, based in San Diego, California, says that it is set

NUMBER CRUNCH

1,004

The number of rhinoceroses illegally killed in 2013 in South Africa, according to government figures. It marks the nation's worst year on record for rhino poaching.

Source: Dept Environmental Affairs

to produce the first platform capable of sequencing an entire human genome for less than US$1,000. Other companies have previously promised to hit that elusive price target, but Illumina's new HiSeq X Ten could come closest. Jay Flatley, the company's chief executive, announced on 14 January that the $10-million sequencing system will be available to customers this year. See go.nature.com/jhslf4 for more.

Cash for clean coal

The US Department of Energy on 15 January formally approved US$1 billion in funding for FutureGen 2.0, a project to demonstrate carbon dioxide capture and sequestration technology in coal-fired electricity generation. Originally proposed in 2003, then cancelled five years later, the FutureGen programme was revived in 2009 by the administration of President Barack Obama. The current plan will retrofit part of a power plant in Meredosia, Illinois, to burn coal under high-oxygen conditions, producing a purified stream of CO_2 emissions that can be stored underground.

Europe's new drugs

The European Medicines Agency (EMA) recommended that 38 drugs with a novel active ingredient be authorized for use in Europe in 2013, compared with 35 in 2012, 25 in 2011 and 15 in 2010. In total, 81 medicines were recommended for human use last year, the London-based EMA said on 20 January. (The European Commission must approve agency recommendations before drugs can be marketed, but it rarely declines such requests).

EVENTS

Flu resurgence

The H7N9 avian influenza virus is making a comeback in China. In the first two weeks of January, 47 people fell ill, the World Health Organization has reported. That compares with more than 60 cases over a comparable period last April, the month that infections peaked. The virus had been relatively quiet since May, following the closure of live bird markets, until December, when a large uptick of 15 cases occurred. Bustling live markets gearing up for the Chinese New Year are thought to have contributed to the latest rise.

Comet craft

JÜRGEN MAI/ESA

The European Space Agency's Rosetta spacecraft has woken up after almost three years of hibernation in space. The craft, part of a €1-billion (US$1.4-billion) mission to track down a comet, was turned off in 2011 to save energy while travelling in deep space. Rosetta successfully re-established communications with Earth on 20 January, to much jubilation at the agency (**pictured**). The spacecraft will now journey to its target, the comet 67P/Churyumov–Gerasimenko, which it will approach in August and observe at close quarters before attempting to release and land a probe in November. See go.nature.com/1igyyz for more.

Journal shut down

The German academic publisher Copernicus Publications announced last week that it had shut down its journal *Pattern Recognition in Physics*, citing what it called nepotistic reviewing that it regarded as malpractice. The company, based in Göttingen, was responding to a special issue on solar variability published late last year, in which the issue's editors doubted global warming. Copernicus will keep all controversial papers online, it said, but now wants to distance itself from "the apparent misuse of the originally agreed aims and scope of the journal". See go.nature.com/pexi6j for more.

Water worries

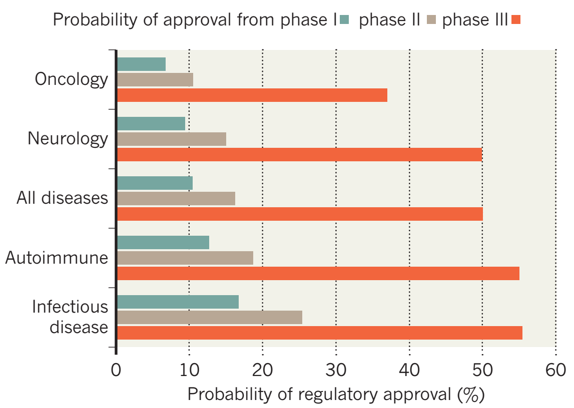
Residents of West Virginia slowly regained access to clean drinking water last week, following the discovery of a chemical spill on 9 January that contaminated the supply for hundreds of thousands of people. The Elk River became tainted when storage tanks at an industrial plant in Charleston, West Virginia, leaked 4-methylcyclohexane methanol, a chemical used to process coal. The spill has prompted the US Senate Environment and Public Works Committee to plan two hearings on chemical safety. The plant's owner, Charleston-based Freedom Industries, filed for bankruptcy on 17 January.

TREND WATCH

Only one in ten pathways for drug development ended in US regulatory approval between 2003 and 2011, according to a recent analysis (M. Hay *et al.* *Nature Biotechnol.* **32,** 40–51; 2014). By separately analysing each disease for which a single drug was tested (see chart), the authors found lower drug-success rates than estimated by many previous studies. Nearly 15% of vaccines entering phase I trials gained eventual approval, whereas small-molecule drugs succeeded at only half that rate.

PHARMA PIPELINE

US approval rates for drugs may be lower than previously estimated, with oncology drugs faring worst.



SOURCE: NATURE BIOTECHNOL.

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

KEVIN LAMARQUE/REUTERS

US President Barack Obama signed a $1.1-trillion budget into law last week.

POLICY

Budget offers recovery hope

US physical sciences benefit more than biomedical research.

BY LAUREN MORELLO, JESSICA MORRISON, SARA REARDON, JEFF TOLLEFSON AND ALEXANDRA WITZE

For US science agencies, the worst may be over. After a year of pitched budget battles, which began with an across-the-board cut and ended with a government shutdown, Congress displayed a rare moment of harmony last week as its members swiftly approved a US$1.1-trillion spending bill for 2014.

The bill, which President Barack Obama signed into law on 17 January, funds government operations until 30 September. It reverses much of the damage wrought in 2013 by the automatic 5.1% cut known as sequestration. And it even puts some agencies above 2012 levels — the year against which agencies are now measuring themselves after the *annus horribilis* of 2013 (see 'Budget highlights'). "These numbers are about as good as anybody could have expected," says Matt Hourihan, director of the research and development budget and policy programme at the American Association for the Advancement of Science in Washington DC. In general, he says, science agencies received bigger boosts than other sectors of the government — a sign, perhaps, that bipartisan support for science endures despite an overwhelming push for fiscal austerity.

BASIC RESEARCH

But that relief was not administered equally within science. Agencies that fund physical-sciences research, including the National Science Foundation (NSF) and the Department of Energy's Office of Science, received small increases compared with 2012 funding levels. Biomedical research, by contrast, was dealt a blow, with the National Institutes of Health (NIH) seeing its budget decline by roughly $800 million from 2012 levels to $29.9 billion. That continues a decline that began in 2004, after adjusting for inflation.

The Food and Drug Administration did not fare much better than the NIH, although its $2.6-billion budget does surpass its 2012 funding level by $45 million.

Passing a budget in the middle of flu season, however, seems to have rustled up money for public health. The $5.8 billion approved for the Centers for Disease Control and Prevention (CDC) adds $151 million to its 2012 budget. It includes $255 million for purchasing vaccines through Project BioShield — $5 million more than the president requested — and $30 million for the Advanced Molecular Detection initiative, which will upgrade the CDC's bioinformatics technology and enable faster detection of outbreaks.

Perhaps the biggest single boost within biodefence is the $404 million for the long-awaited, controversial National Bio and Agro-Defense Facility (NBAF) in Manhattan, Kansas, which will study infectious diseases in large livestock. The Department of Homeland Security project was delayed by Congress after a 2010 National Academy of Sciences report found faults in the facility's design that could allow pathogens to escape.

Although the NBAF funding falls $310 million short of Obama's request, Ron Trewyn, vice-president for research at Kansas State University in Manhattan, says that construction of the facility could start as early as this year. The lab, which is expected to take five years to complete, will be the only US livestock-disease facility to have the highest level of biosecurity (known as BSL-4).

The NSF received $7.2 billion, an increase of $70 million over 2012 levels. Its budget includes $200 million for ongoing construction projects, and sets aside $17.5 million for the Large Synoptic Survey Telescope (LSST), a $653-million astronomy project in Chile. Although the sum for the project is $10 million short of the president's request, Congress has essentially blessed the LSST, and construction can now begin in July. The fact that the telescope was explicitly mentioned in the bill left some of its proponents pleasantly surprised. "Getting a new start in an NSF budget is a very major accomplishment," says William Smith, president of the Association of Universities for Research in Astronomy in Washington DC, which will manage the LSST project for the NSF.

Meanwhile, NSF-funded political scientists are equally excited not to get a mention. In the 2013 spending bill, Senator Tom Coburn (Republican, Oklahoma) inserted a provision requiring that political-science research funded by the NSF benefit national security or US economic interests. The NSF then skipped a round of grants in August. The absence of restrictions in the latest bill "returns the judgement of scientific merit to those who are in the best position to gauge the quality of the work", says Rick Wilson, a political scientist at Rice University in Houston, Texas, and a former programme director for political science at the NSF.

ENERGY

At the Department of Energy, the bill provides $27.3 billion — $1.7 billion below the president's request but still nearly $1 billion above the 2012 level. That includes nearly $5.1 billion for the Office of Science, an increase of 3% from 2012. Within the office, researchers in fusion energy received one of the biggest surprises. The budget boosts funding for their field by 28% compared with 2012 levels, providing nearly $306 million for the domestic fusion research programme. It includes $22 million to revive Alcator C-Mod, a magnetic plasma-confinement machine at the Massachusetts Institute of Technology in Cambridge that the administration sought to cancel in 2012. The bill also provides the full $200 million US contribution to ITER, an international fusion experiment under construction in southern France.

BUDGET HIGHLIGHTS

How science agencies fared in the budget (US$ millions).

Agency	2012	2013*	2014
National Institutes of Health	30,702	28,926	29,926
Centers for Disease Control and Prevention	5,656	5,437	5,807
Food and Drug Administration	2,507	2,386	2,637†
National Science Foundation	7,105	6,884	7,172
NASA (science)	5,074	4,782	5,151
Department of Energy's Office of Science	4,934	4,621	5,071
National Institute of Standards and Technology	751	769	850
Environmental Protection Agency	8,449	7,918	8,200
National Oceanic and Atmospheric Administration	4,906	4,740	5,315
US Geological Survey	1,068	1,021	1,032

*2013 figures include the roughly 5% across-the-board cut arising from the sequester. †Includes one-time transfer of $85 million in user fees.
Source: US Congress; White House Office of Management and Budget.

The budget represents a setback for the president's clean-energy research agenda, however. Although overall spending on renewable energy and energy efficiency is nearly 7% higher than in 2012, the budget eliminates much of the extra funding requested by the president for clean-energy research and development, in areas ranging from advanced manufacturing to wind and solar energy. The plan also cuts $99 million from the president's request for the Advanced Research Projects Agency-Energy, and so its budget will remain relatively flat at $280 million. Fossil-fuel research and development, meanwhile, will receive $562 million, which is $142 million above the president's request and $225 million above 2012 levels — an increase of 67%.

"These numbers are about as good as anybody could have expected."

SPACE

NASA fared reasonably well, receiving $17.6 billion overall, including almost $5.2 billion for science. However, lawmakers took a public swipe at the proposed mission to bring an asteroid to lunar orbit and send astronauts there to study it (see *Nature* **499,** 261–262; 2013). "NASA has not provided Congress with satisfactory justification materials such as detailed cost estimates or impacts to ongoing missions," they wrote. Although the agency is hunting for more asteroids of the right size and orbit to be a mission target, the list of candidate rocks is just six — down from 14 last summer, because some have been ruled out as being too small.

The budget also carries language that bans NASA from engaging bilaterally with China. Originally introduced by Congressman Frank Wolf (Republican, Virginia) in 2011, the China restrictions are politically popular among Republicans in the House of Representatives and are likely to be kept in place by Wolf's successor after he retires this year, if Republicans keep control of the House after the 2014 elections, says Marcia Smith, a space-programme analyst and founder of SpacePolicyOnline.com. The ban has led to much confusion about how widely it applies. For example, Chinese scientists were initially barred from a November 2013 exoplanet conference at a NASA facility in California before later being allowed in.

From the $1.3 billion specified for planetary exploration, Congress required that $80 million would go to prepare a mission to Jupiter's moon Europa, even though a similar mission was dropped from a joint plan to explore the Jupiter system with the European Space Agency. The fact that some members of Congress specifically fund a Europa mission each year has rankled with planetary scientists who want to study other parts of the Solar System. Lawmakers did, however, encourage NASA to put out a call for ideas by May for its next Discovery class of small-scale planetary missions.

LOOKING AHEAD

Speculation now turns to the 2015 budget cycle, which begins on 1 October. For the past few years, Congress has bridged its political differences on government spending by approving a series of stopgap measures that maintained prior funding levels, avoiding major changes to agency budgets. Now, buoyed by the new deal, House and Senate leaders say that they hope to enact a detailed funding plan before the October deadline. Lawmakers have already overcome one major obstacle by setting an overall 2015 level for 'discretionary' spending, a category that includes civilian-science agencies. To Hourihan, it is reason for cautious optimism. "Things may get a little bit easier," he says. "But the big picture hasn't really changed." ■

PETE BUCKTROUT/BAS

A 2012 attempt to drill through a 3-kilometre-thick Antarctic ice sheet faced technical difficulties.

ANTARCTIC RESEARCH

Polar drilling problems revealed

Report into failings of expedition to explore Antarctic lake finds equipment to blame — but complications can be fixed.

BY QUIRIN SCHIERMEIER

Christmas Day 2012 was a very bad day for glaciologist Martin Siegert and his team of Antarctic researchers. After weeks of equipment failures, Siegert was forced to halt an ambitious attempt to drill into a lake deep beneath the West Antarctic Ice Sheet. "The decision was difficult to make but easy enough to call," he noted in his field diary. Exhausted and disappointed, the team packed up.

But it has not given up. Over the past year, researchers, engineers and officials involved in the US$12-million drilling project, funded by the UK Natural Environment Research Council, have carried out and responded to several internal reviews into the reasons for its failure. And now, in a paper under review by the *Annals of Glaciology*, Siegert, who is based at the University of Bristol, UK, and his colleagues have summarized the problems that they suffered at Lake Ellsworth and laid out options for putting them right. They think that several years of engineering work will be required to develop improved technology for a more reliable drill, but they say that success is achievable.

"I am glad to see that they plan to publish their drilling efforts," says John Priscu, a glaciologist at Montana State University in Bozeman who has worked on similar lake-drilling projects in western Antarctica. "It will quell rumours and provide a solid bit of groundwork on which they can move forward."

Lake Ellsworth is one of several hundred lakes beneath Antarctica's ice sheets (see 'Hidden lakes'). Scientists suspect that the extended basins, isolated for possibly millions of years, support specially adapted forms of life. Organisms that may thrive in the extreme environment could even bear clues as to the biology of extraterrestrial life, such as any that might exist in a suspected ocean beneath the icy surface of Europa, one of Jupiter's moons.

The Ellsworth project was in planning for more than ten years. The teardrop-shaped lake, about 15 kilometres long and up to 156 metres deep, and in a valley, was extensively charted with seismic methods and ice-penetrating radar by the team before the drilling attempt.

After arriving at the lake in early December 2012, Siegert's team had hoped to cut through the 3-kilometre-thick ice sheet in a single 72-hour effort. A specially developed hot-water drilling technique, devised by engineers at the British Antarctic Survey, was designed to minimize air and water pollution.

According to the paper, problems started when the boiler that was intended to melt large quantities of snow to provide hot water for the drill failed to work properly because of short-circuiting in its control panel. More severe problems followed. The two parallel drills — one to drill the main borehole to reach the lake, and one to create a reservoir cavity to recirculate drilling water — ran too slowly. Other failures, including of components designed to ensure vertical drilling, exacerbated the problems.

"The drilling was essentially undertaken blindly," says Siegert. Probably because one or both holes were not drilled vertically, the cavity failed to link with the main borehole. Water also leaked into the cavity drill and froze the hose in the drill hole. Attempts to remove the hose failed, so it had to be cut. At that point, and with not enough fuel left to reach the lake, Siegert gave up.

His report into the project suggests a number of steps to improve the drilling system: sensors need to be thoroughly tested for reliability under conditions comparable to Antarctica's harsh environment, spare parts must be available on the site, and a field team must include electrical engineers to assist with on-site operations.

The method is not fundamentally flawed, however. Last year, Priscu and his team successfully used a hot-water drill to explore Lake Whillans, a small body of water on the edge of the Ross Ice Shelf in western Antarctica. The lake, which is not as deep as Ellsworth, does seem to harbour microbial life (see *Nature* http://doi.org/q3m; 2013).

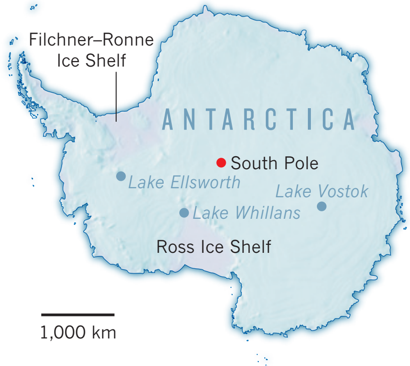
There are other reasons for pursuing hot-water drill technology. In 2012, Russian scientists broke into Lake Vostok, by far the largest of Antarctica's hidden lakes, using a kerosene-fuelled drill. But their samples are spoiled with drill fluid and the bacteria they contain are probably contaminant species.

Consideration of what went wrong at Ellsworth should result in a revised plan for a return mission, says Mahlon Kennicutt, chair of the Lake Ellsworth advisory committee. The team hopes to formally propose a second attempt in five years.

"Antarctic science is by its nature risky," says Kennicutt. "However, the potential gains in knowledge outweigh the costs and the risks in most cases, and this is especially true for the exploration of subglacial aquatic environments." ■

HIDDEN LAKES

Antarctica has about 400 subglacial lakes, some of which have been isolated for millions of years.



SIIM SEPP/ALAMY

Estimates suggest that olivine could be used to sequester a significant proportion of carbon emissions.

GEOCHEMISTRY

Rock's power to mop up carbon revisited

Experts push for more research into olivine weathering.

BY DANIEL CRESSEY

Last week, a group of geoengineers met in Hamburg to discuss what on the face of it sounds like a very attractive idea: to soak up anthropogenic carbon emissions using only rocks and water. In particular, they want to help to mitigate climate change by crushing rocks and dropping them into the sea or spreading them on land. The meeting was hailed a success, but the idea is still far from fruition.

The 'weathering', or breaking down, of rocks is a hugely important but very slow part of the carbon cycle. Natural weathering locks up atmospheric carbon dioxide by means of chemical reactions between common silicate minerals and air. For example, when magnesium-rich olivine, a rock of particular interest to geoengineers, is brought together with CO_2 and water under natural conditions, the resulting reaction forms magnesium carbonate and silicic acid, thereby removing and storing carbon.

But some scientists think that this natural process could be exploited to offset at least some of the carbon emitted by human activities. Rather than waiting for rocks to be slowly weathered away, olivine could be mined on an industrial scale, ground up, and spread over land or in the sea, speeding up these chemical reactions and sucking vast quantities of CO_2 out of the atmosphere. But this presents practical problems: according to one estimate, you would need to spread 5 gigatonnes of olivine on beaches annually to offset 30% of global CO_2 emissions (assuming 1990 levels of emissions; S. J. T. Hangx & C. J. Spiers *Int. J. Greenhouse Gas Contr.* **3,** 757–767; 2009).

At the informal meeting, about 20 enhanced-weathering experts discussed recent research in the area and tried to summarize and coordinate future work, for example by agreeing to standardize experiments. Until now, there has been no organized research agenda for the fledgling field, says meeting convener Jens Hartmann, who works on geological cycles and carbon sequestration at the University of Hamburg in Germany. "It was very positive; we know we are now a community," he says.

> ***"We have good and very promising results, but there are still a lot of unknowns."***

Hartmann points out that humans have been exploiting rock weathering for decades — for example, by spreading minerals such as olivine, pyroxenes and serpentines as fertilizers. "The question is, can we optimize it and can we do it in areas we are not doing it?" he says.

As with its use as a fertilizer, olivine would have to be finely crushed to maximize its exposure to carbon. Olaf Schuiling, a geochemist at Utrecht University in the Netherlands and a passionate advocate of enhanced weathering, proposes spreading coarse olivine grains on beaches that experience heavy seas. "There the grains are tumbling around in the surf and the waves, they collide, they abrade each other, and produce very rapidly a lot of tiny olivine slivers that weather quickly," he says.

However, there is little evidence for the practical rates of weathering that could be expected if large amounts of olivine or other rocks were mined and spread on fields or dumped into the sea. This, in turn, means it is not clear how much would be needed to significantly mitigate carbon emissions, how long it would take to work or whether it would be cost and energy efficient.

In theory, one kilogram of olivine sequesters about one kilogram of CO_2, but the rate at which this happens can be slow. And the actual efficiency of sequestration will be much lower than 100%, because of the energy used — and emissions released — in grinding and transporting the rock. In some cases, this could emit more carbon than would be sequestered.

Francesc Montserrat, a marine benthic ecologist at the Royal Netherlands Institute for Sea Research in Yerseke, is trying to pin down the figures. He is using small tanks to measure the weathering of olivine in various conditions — including the impact of worms that live in and eat the sandy sediment. Montserrat's experiments will test the idea that when these worms eat tiny grains of olivine they also help to break down the crust that can form on olivine's surface, which slows down the weathering effect.

"You need to have some hard numbers to go to the authorities to say whether it will be safe enough to try it out," he says. "We have good and very promising results, but there are still a lot of unknowns."

Even advocates of this method of geoengineering admit that large-scale enhanced weathering is not without risk. Olivine can contain toxic heavy metals such as nickel that could accumulate in the environment. Grinding rocks would produce dust, which might harm human health. And putting olivine into the sea could change the pH of the water, helping to combat ocean acidification driven by climate change but also potentially harming marine organisms by altering their environment.

Phil Renforth studies carbon sequestration and minerals at the University of Oxford, UK, and attended the Hamburg meeting. He says that there is a pressing need to conduct more work on enhanced weathering given that carbon emissions are likely to continue to rise, and because of the current focus on dealing with emissions by capturing them from power stations and storing them underground.

"We're putting all our eggs in one basket if we're only looking at one method," he says. There's a real need to diversify the portfolio." ■

GEORGE SANKER/NATURE PICTURE LIBRARY

Grizzly bears in Yellowstone National Park are switching from eating pine nuts to eating meat as global warming has allowed pests to kill pine trees.

ECOLOGY

Yellowstone grizzlies face losing protected status

Conservationists protest after panel recommends ending bears' endangered-species listing.

BY LAUREN MORELLO

For the US government, the grizzly bears of Yellowstone National Park in Wyoming embody a stunning success story: a population resurgent after 40 years of protection under the Endangered Species Act. More than 700 bears now roam the region, up from 136 in 1975, when the grizzly (*Ursos arctos horribilis*) was listed as threatened after decades of deadly clashes with ranchers, hunters and park visitors. But the US Fish and Wildlife Service is now expected to lift the legal safeguards, after a government advisory panel of wildlife officials endorsed delisting the bear last month.

Conservation groups have pushed back, saying that the government has underestimated the threat that climate change poses to the bears' food supply, especially stands of whitebark pine. As the Yellowstone region has warmed, mountain pine beetles and blister rust fungus — once thwarted by the cold, dry climate — have devastated the trees, depriving grizzlies of energy-rich pine nuts. Moreover, say conservationists, invasive fish have crowded out native cutthroat trout in Yellowstone Lake at the heart of the park, reducing another important food source for the bears.

"We have an unprecedented situation with deteriorating foods, and an ecosystem that is unravelling," says Louisa Willcox, the Northern Rockies representative at the Center for Biological Diversity in Livingston, Montana. The centre was one of several groups that sued the US government in 2007, following an earlier attempt to delist the bear. After two years, a district-court judge restored protection, citing concerns about the declining whitebark pine and its effect on the bears' diet.

A report delivered in November by the US Geological Survey's Interagency Grizzly Bear Study Team describes a resilient and healthy bear population that has adapted to the loss of pine nuts by eating more elk and bison, keeping fat stores at levels that allow the bears to survive and reproduce. For Christopher Servheen, a biologist who oversees grizzly-bear recovery efforts at the Fish and Wildlife Service in Missoula, Montana, that is not surprising. "Bears are flexible," he says. "It's easier to say what they don't eat than what they do eat."

But other researchers suspect that the change carries a steep price. "Eating meat is hazardous on all fronts," says David Mattson, an ecologist at Yale University in New Haven, Connecticut. A reliance on meat heightens

HOME ON THE RANGE

A growing Yellowstone grizzly-bear population now extends beyond the national park.

SOURCE: IGBST

the risk that adult bears will come into contact with humans, including livestock owners and hunters seeking elk, he says. For young bears, it may increase the frequency of potentially deadly interactions with aggressive adult male bears and wolves.

Critics also argue that the government is basing its decisions on flawed population estimates. A study published last July suggests that the government's figure of 741 bears is inflated (D. F. Doak and K. Cutler *Conserv. Lett.* http://doi.org/q3d; 2013). The number of survey flights used to count bears has tripled since the mid-1990s, but, the study argues, the model used to extrapolate population figures from the flights' tallies does not account for increased observation time. Further distortion may arise because the model assumes that female bears will reproduce consistently throughout their 30-year lives, with no decrease in fertility as they age.

Mattson says that population estimates have in the past jumped by more than 100 bears when the statistical method has shifted. "There is no clean and simple way to estimate the size and trend of the Yellowstone population," he says.

But those criticisms are rejected by Frank van Manen, a wildlife biologist with the US Geological Survey in Bozeman, Montana, who led the diet study. Observation time has increased, he says, but so has the grizzly bears' range (see 'Home on the range'), which cancels out any observer bias from increased search hours. And although the government's official estimate of the population did jump from 629 to 741 bears this year, van Manen says that the new number is better. That is in part because the revision takes into account a 2011 demographic study of bear survival rates based on radio-collar tracking data — the first such study since 2002 — that gives biologists more confidence in their population surveys.

Servheen says that if the government were to decide to pursue delisting, as many expect, the decision would not be announced until late spring at the earliest. At that point, the Fish and Wildlife Service would open a 60-day public-comment period to seek reaction.

But even that is unlikely to be the last word on the grizzlies: conservation groups are already gearing up to sue. Perhaps the only point on which the US government and its opponents agree is that there will be more legal wrangling over the Yellowstone bears' future. "It's sad that it's come to this," says Servheen. "What it should be is a celebration." ■

EARTH SCIENCE

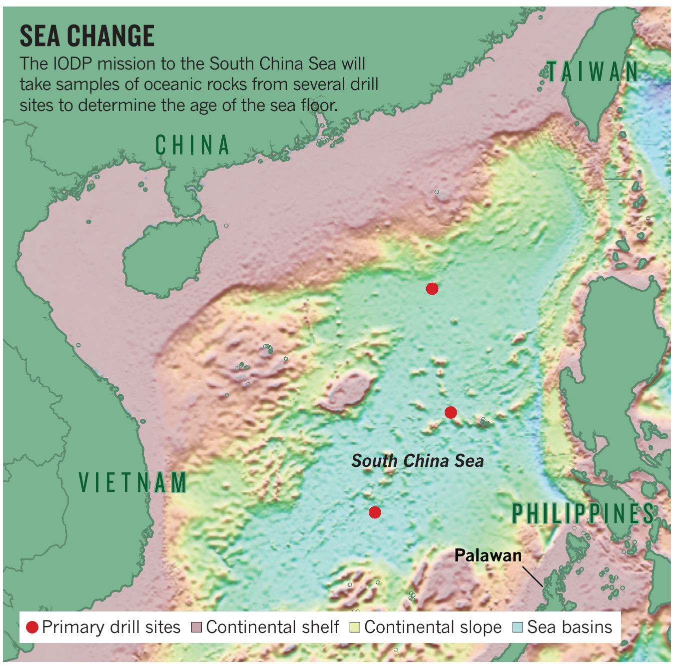
Sea drilling project launches

International expedition hopes to unravel mysteries of the South China Sea, one of the world's most geologically important seas.

BY JANE QIU

The South China Sea is well known for its geopolitical tensions, but less is known about its many geological stresses and strains. That is set to change.

On 28 January, an international team of scientists — from countries including China, the Philippines, India and the United States — is due to set sail from Hong Kong on board the research vessel *JOIDES Resolution*, marking the first expedition of the International Ocean Discovery Program (IODP), formerly known as the Integrated Ocean Drilling Program. Its aim is to determine the age of the South China Sea, and to resolve ongoing controversy over how it formed.



CHUN-FENG LI

With an area of more than 3 million square kilometres and thousands of islands and reefs, the sea occupies a scientifically interesting position between the world's highest mountains, the Himalayas, and the deepest point on Earth's surface, the Mariana Trench in the western Pacific Ocean.

It is "a natural laboratory for studying continent break-ups and sedimentary-basin formation", says Dieter Franke, a geologist at the Federal Institute for Geosciences and Natural Resources in Hannover, Germany, who is not involved in the expedition. The sea's relatively small size and young age (between 25 million and 42 million years old) compared with major ocean basins (the Pacific plate can be traced back at least 200 million years) mean that it is possible to probe its entire history through just a couple of IODP expeditions, Franke says.

Little is known about the formation of the South China Sea. The crust beneath it was created after a part of Eurasia that once stood in its place began to stretch in a north–south direction. As the stretching continued, the continent became progressively thinner. At some point it broke apart, releasing magma that solidified and moved away from the eruption sites — a process called sea-floor spreading. The land mass drifted south, breaking into pieces and giving rise to islands such as Palawan in the Philippines and Borneo.

But for decades, geologists have been debating what triggered the continent to stretch and break up in the first place. Among the ideas proposed are that it was caused by the collision between Eurasia and the Indian subcontinent; that the continent buckled as an ancient oceanic plate slid beneath present-day southern China; or that the Pacific plate pulled away from the the Eurasian coast.

"The hypotheses are based only on circumstantial evidence," says Jian Lin, a marine geophysicist at the Woods Hole Oceanographic Institution in Massachusetts and co-chief scientist on the drilling project. "Much of the controversy stems from different estimates of how old the sea floor is." Until recently, scientists have had to make age estimates by towing

PER-ANDRE HOFFMANN/PICTURE PRESS/ALAMY

The break-up of Eurasia tens of millions of years ago led to the formation of islands such as Palawan.

a device called a magnetometer along the sea surface. Oceanic rocks capture the direction of Earth's magnetic field at the time they formed, and this information can be used to date them.

But "there are different ways to interpret the data, and the results can vary wildly", says Paul Tapponnier, a geologist at Nanyang Technological University in Singapore, who is not involved in the trip. The sea-floor spreading of the southwest sub-basin, for instance, is thought to have begun between 25 million and 42 million years ago, and to have ended between 16 million and 35 million years ago. "The only way to resolve the controversy is to measure the age of the ocean crust directly," he says.

Over the next two months, the team will drill up to 2 kilometres into the seabed to collect rock samples (see 'Sea change'). Geochemical and geophysical analyses will then allow the researchers to determine the rocks' ages and characteristics. These should yield clues to their origins. By drilling at different sites, the scientists should be able to tell precisely when the sea floor started to spread and when the process ended.

"This is a fundamental question that needs to be addressed before we could even begin to piece the puzzle together," says Chun-Feng Li, a marine geophysicist at Tongji University in Shanghai, China, and the other co-chief scientist on the project. Once the precise age of the sea floor is known, researchers will be able to make correlations with the timing of other events associated with the creation of the South China Sea, such as the retreat of the Palaeo-Pacific plate. Identifying the origin of the rocks will also help to pinpoint which of the continental break-up hypotheses is the most likely.

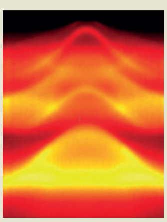
The expedition "will be just the first step towards a comprehensive understanding of how the South China Sea and other marginal seas opened", says Li. A plan for a follow-up project to investigate the rifting process in more detail has already been submitted to the IODP.

The importance of unravelling the geological history of the South China Sea "goes beyond academic curiosity", says Franke. Oil and gas normally accumulate at the continental margins at which rifting takes place, and a better understanding of when and how the basins formed will help to locate new reserves, he says. It could also facilitate earthquake research of the Manila Trench in the Pacific Ocean, says Alyssa Peleo-Alampay, a marine geologist at the University of the Philippines Diliman in Quezon City and a project member. The trench came into being as the oceanic crust of the South China Sea began to sink beneath the Philippine Sea plate — a process that continues today and causes frequent quakes. "A proper understanding of the South China Sea is long overdue," she says. ■

MORE ONLINE

SCIENCE/AAAS

TOP STORY



'Dirac semi-metals' are three-dimensional analogues of graphene go.nature.com/73elun

MORE NEWS

- El Niño oscillations may become more frequent go.nature.com/haig5d
- Quasar light reveals possible intergalactic filament go.nature.com/rqxu1b
- Solvent breaks down biomass without enzymes go.nature.com/l7gohb

CLONING COMEBACK

Ten years ago, Woo Suk Hwang rose to the top of his field before fraud and dodgy bioethical practices derailed his career. Can a scientific pariah redeem himself?

BY DAVID CYRANOSKI

The Sooam Biotech Research Foundation nestles on a wooded hillside in Guro, a district on the southwestern outskirts of Seoul. Spartan, quiet and cold on this winter day, the grey-white exterior belies the buzz of activity within.

A door just off the foyer leads to a corridor of canine chaos. In stalls to the left, Tibetan mastiff and Australian shepherd puppies are cavorting. A Yorkshire terrier dances back and forth on its hind legs. And an adult mongrel howls with separation anxiety, only calming down when the two beagle pups that she gave birth to are returned to her pen. She doesn't know that she is just a surrogate mother, nor that the pups are highly unusual dog clones, engineered to show the symptoms of Alzheimer's disease.

The right side of the corridor houses a wall-sized window that looks onto an operating theatre. Inside, Woo Suk Hwang, in a blue surgeon's gown, cap and mask, is working on a bitch in labour. He greets his visitors through a microphone headset and then explains that this is an emergency: one of the puppies is stuck in the cervix. He makes an incision and carefully probes the dog's womb until the whitish sausage of a puppy emerges. After it is wiped down, Hwang holds it to his ear, listening for sounds of breathing. He then gently massages the groggy pup into consciousness and goes back for the last one. Minutes later he announces: "We have saved all three cloned dogs." Hwang brims with pride.

Eight years ago, few could have imagined watching such a jubilant scene. Hwang, a world-famous cloning researcher, had just plummeted from the pinnacle of scientific success, when it became clear that he had committed fraud in two articles[1,2] describing stem-cell lines derived from cloned human embryos. There had been gross ethical lapses in the way Hwang had collected the human eggs for his experiments, and the papers were found to contain fabricated data. They were eventually retracted. It was one of the most widely reported and universally disappointing cases of scientific fraud in history. In January 2006, Un-chan Chung, then president of Seoul National University (SNU), where Hwang had done the work, called the episode "an unwashable blemish on the whole scientific community as well as our country".

If the stain cannot be washed away, perhaps it can be stamped out of memory by hundreds of paws and hooves. With private funding from steadfast fans, Hwang opened Sooam in July 2006. He has since cloned hundreds of animals — dogs, cows, pigs and coyotes. His goals include producing drugs, curing diabetes and Alzheimer's disease, providing transplantable organs, saving endangered species and relieving grief-stricken pet owners. He has a raft of publications in respectable journals, collaborations within and outside South Korea, and increasing institutional support from government agencies. It is hard to square this image with the pictures of Hwang released by the South Korean media in 2005. Shattered by the controversy, he was photographed in a hospital bed, unshaven and reportedly suffering from exhaustion.

Today Hwang plays down his involvement in the fraud. He retains a base of ardent supporters, mostly in South Korea. And he maintains, contrary to scientific consensus, that he really did create the first line of cloned human embryonic stem cells. He has even had success in getting some legal recognition of that claim.

In December, he welcomed reporters into Sooam to tour the facilities and see him deliver some cloned puppies, but he declined to comment for this story. Maybe "in a couple of decades", he wrote by e-mail.

CLONING FOR COUNTRY

A veterinarian by training, Hwang rose to fame in South Korea in the late 1990s by cloning animals — and by developing important allies (see 'The rise and fall and rise of Woo Suk Hwang'). He asked then-President Kim Dae-jung to name the first cloned beef bull, and he promised a national agricultural boom centred on cloned cattle.

His popularity in South Korea grew, and in 2004 he shot to international fame when *Science* published a paper[1] in which he claimed to have created an embryonic-stem-cell line from a cloned human embryo — something that several groups had been trying to do. Hwang's success seemed to offer an endless supply of versatile cells genetically matched to the cell donor. Through this process, often called therapeutic cloning, it was hoped that doctors could rejuvenate failing tissues or organs, or that cells derived from people with virtually any disease could be used for research and drug screening. The following year, his group published a second paper[2], describing the development of 11 more such lines, making the process so routine that clinical application seemed imminent.

LEFT: AHN YOUNG-JOON/AP; RIGHT: DAVID CYRANOSKI

Snuppy, the first cloned dog (left), was one of Woo Suk Hwang's successes. Today (right) Hwang regularly delivers cloned animals at an institute near Seoul.

But even as his star was rising, cracks were beginning to show. In May 2004, one of Hwang's graduate students told *Nature* that she had donated eggs for experiments in the first paper (see *Nature* **429,** 3; 2004). It was a controversial assertion: many bioethicists worry that, in such a situation, students might feel pressure to endure a risky and uncomfortable procedure.

Hwang denied the charge and the student recanted her statement. But in November 2005, amid increasing evidence, Hwang admitted that he had lied (see *Nature* **438,** 536–537; 2005). Two students had donated eggs; Hwang even drove one to the clinic, where she donated her eggs before returning to the lab to try to make cell-line clones of herself. Hwang had also paid donors for eggs used in the 2004 paper, contradicting what the paper said. And he continued to compensate donors even after a South Korean bioethics law came into effect in January 2005 banning the practice.

Hwang's triumphs soon unravelled further. In January 2006, an SNU investigation committee announced that both of his human-cloning papers were fraudulent. The committee found that the cell line reported in 2004, called NT-1, was not produced by cloning and was probably a product of parthenogenesis — the 'virgin birth' process by which an egg starts embryonic development without the contribution of sperm. The 11 stem-cell lines claimed to be patient-specific clones in the 2005 paper turned out to be normal embryonic-stem-cell lines from a fertility hospital that had been relabelled. Images and graphs in both papers were fabricated to give the appearance of clones. "The research team of Professor Hwang does not possess patient-specific stem cell lines or any scientific bases for claiming having created one," the report concluded.

Hwang's empire crumbled. He was expelled from SNU in March 2006. The Seoul Prosecutor's Office raided his laboratory and launched a massive investigation.

Hwang took responsibility for poor oversight of his lab, but maintained that he had been duped by a co-author. During the investigation, one co-author admitted to switching stem cells without Hwang's knowledge, but Hwang also admitted to ordering subordinates to fabricate data. A complicated web of blame emerged in which Hwang admitted being involved in fraud but still maintained that the achievement was real.

> THE EPISODE DREW ATTENTION AND INTEREST FROM GOVERNMENT AND ORDINARY PEOPLE.

Data fabrication is not illegal in South Korea, but knowingly using bogus articles to get funding is. The Prosecutor's Office charged Hwang with fraud, embezzlement and bioethics violations, and a three-year court case ensued. In 2009, the court threw out the fraud charge, saying that the companies involved gave the money knowing that they would not benefit from the donation. Hwang was, however, convicted of violating the country's bioethics law and of embezzling government funding. He was sentenced to two years in prison. The term, later reduced to 18 months, is still under appeal in court. But even if Hwang loses his appeal, as long as he doesn't break the law during his probation, he will not spend any time in jail, says Sean Hayes, a partner at IPG Legal in Seoul.

DOGGED PURSUIT

Despite his legal troubles — and the widespread belief that his career was over — Hwang continued to work, thanks to the supporters who amassed US$3.5 million to launch Sooam. About 15 scientists followed Hwang from SNU, and around half of those remain today among Sooam's 45 staff. His team now creates some 300 cow and pig embryos per day, and delivers about 15 cloned puppies per month.

Hwang has long been interested in cloning dogs. He reported[3] the world's first cloned puppy in 2005 — a claim upheld by the SNU investigation. Since 2006, Sooam has cloned more than 400 dogs, mostly pets. Customers, the majority of whom are from the United States, pay about US$100,000 for the service. Sooam has begun supplying dogs to the Korean National Police Agency in Seoul in the hope that clones of proven service animals will quickly learn their trade as sniffer dogs. And last year, it launched a contest for a UK dog owner to have a dog cloned for free — which would make it the first cloned canine in the country.

Although Sooam could make more money from cloning pets if it cut prices and increased production, the non-profit organization wants to be more than a dog-cloning factory. "It's just a

THE RISE AND FALL AND RISE OF WOO SUK HWANG

FEBRUARY 2004
Woo Suk Hwang describes the first stem-cell line, NT-1, derived from a cloned human embryo.

MAY 2005
Hwang's group publishes a second paper reporting 11 further human embryonic cell lines.

AUGUST 2005
Hwang's group is the first to clone a dog.

NOVEMBER 2005
US collaborator Gerald Schatten splits with Hwang, citing ethical problems in getting human eggs.

DECEMBER 2005
Pushed by increasing evidence, Seoul National University (SNU) launches an investigation.

JANUARY 2006
Hwang's human-cloning research is deemed fraudulent by SNU. His dog-cloning claims are upheld.

FROM LEFT: HWANG WOO-SUK, SNU/AP; LEE JAE WON/REUTERS; KIM KYUNG-HOON/REUTERS; WOO SUK HWANG/SNU/UPI PHOTO/NEWSCOM; SHIN YOUNG-KEUN, YONHAP/AP

side project to get research funding for our other projects," says Insung Hwang, a scientist at the institute who agreed to speak about research at Sooam. He is no relation to Woo Suk.

Using cloning technology, Sooam is creating cows that produce the human interferon protein, which can be used for treating a number of human diseases, in their milk[4], and pigs that are genetically tweaked so that their organs might be suitable for transplantation into humans[5]. Sooam researchers have also created new models for diabetes by putting genes that cause symptoms of the disease in mice into cloned pigs[6] and dogs[7]. Likewise, says Insung Hwang, a transgenic beagle at Sooam that carries a gene related to Alzheimer's disease shows hallmarks of the disease. Researchers at the institute have cloned this beagle 18 more times and are waiting to see whether these dogs also develop the symptoms.

Sooam's ambitions don't stop there. In March 2012, the centre began a collaboration with the Institute of Applied Ecology of the North, part of the North-Eastern Federal University in Yakutsk, Russia. They have joined forces to try to clone a mammoth from ancient tissue dug from permafrost. The project has received great fanfare, but Insung Hwang admits that it is a long shot. "The chances are very small," he says.

Sooam is also expanding its repertoire of species. It has already cloned coyotes (*Canis latrans*)[8] using dog eggs and dog surrogates, and it now hopes to build on that work to clone the African wild dog (*Lycaon pictus*), one of the most endangered carnivores in Africa.

Under Woo Suk Hwang's guidance, the institute has published more than 40 papers documenting cloning successes and technical improvements to the cloning process. "His group is making important yet incremental progress towards long-term goals," says Cindy Tian, a cloning and reproductive biology researcher at the University of Connecticut in Storrs.

The fact that Hwang is being published in peer-reviewed journals is a sign that he is becoming accepted once more. Insung Hwang says that researchers he meets often bring up the fraud and "some reviewers are a little hesitant" to take Sooam manuscripts seriously, but overall, they are treated fairly. Tian, who edited two of Woo Suk Hwang's papers for *PLoS ONE*, says that his "designs are sound and the conclusions are supported with good data". She adds that "it is very unlikely a 'come-back fraudster' would do the same trick again", and that because Sooam work is likely to be closely scrutinized, the researchers there are bound to be on their best behaviour.

Woo Suk Hwang's greatest coup in terms of regaining legitimacy was establishing a partnership in March 2013 with BGI in Shenzhen, China — the world's largest sequencing facility and a powerhouse in scientific publishing (see *Nature* **464,** 22–24; 2010). Together, they plan to look at modifications of chromosomes that determine how genes are expressed, a field called epigenetics. Analysing the variation between clones and how that may contribute to, for example, different coat patterns in dogs could be a powerful tool for such work.

AT SOME POINT, IT WAS CLEAR THAT THE STAKES WERE TOO HIGH FOR DR HWANG TO FAIL.

Yang Huanming, BGI's co-founder, says that he was impressed by the level of involvement from Woo Suk Hwang after watching him deliver a litter of cloned pups. "Personally, I like him, how hard he works, and how passionate he is for science," Yang says.

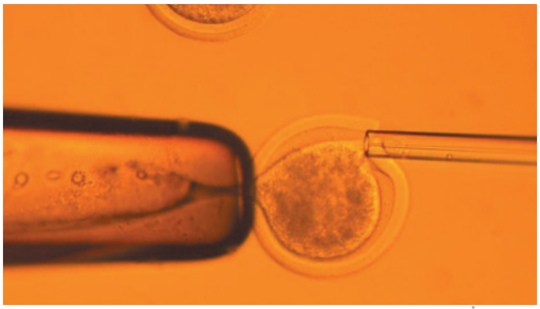
Woo Suk Hwang has also earned support from the Korean government. Roughly 50% of the funding for Sooam now comes from government grants, which includes 3 billion won (US$2.8 million) over three years from Gyeonggi province, Seoul's neighbour, for two cow-cloning projects, according to Insung Hwang. In 2012 and 2013, the Rural Development Administration contributed nearly 190 million won for the interferon project and 140 million won for transgenic animal models of metabolic disease.

But some scientists remain wary. "If you fabricated data once, how would one know that you will not do it again?" asks Hans Schöler, a stem-cell biologist at the Max Planck Institute for Molecular Biomedicine in Münster, Germany. Looking at the unlikely bid to clone a mammoth, Jeong-Sun Seo, director of the Genomic Medicine Institute at SNU, feels a sense of déjà vu. "I am afraid that it seems to be just show," he says. Seo says that he is not opposed to Woo Suk Hwang getting grants for animal cloning, but he draws the line at research into human cloning. Hwang "doesn't know the trends in stem cells. He should stick to his strong animal-cloning technology," Seo says.

LINE OF INQUIRY

Nevertheless, Woo Suk Hwang intends to return to human therapeutic cloning. But he may be trying to ride a wave that has already passed. A competing technology — induced pluripotency, discovered in 2006 — creates stem cells from adult cells, skirting the difficulty of sourcing human eggs and the controversy of embryo destruction. Even the announcement[9] last year that a human stem-cell line had finally been created from a cloned embryo got a more muted reception than the carnival that greeted Hwang when he announced his now-discredited paper.

In 2007, the Korean health ministry gave Sooam approval to do research using human embryos. However, approval to start specific human therapeutic cloning projects has so far been denied twice. Insung Hwang says that no explanation was given, but he thinks that



JULY 2006
Sooam Foundation starts up, with US$3.5 million from Hwang's supporters.

2007
The Korean health ministry grants Sooam the right to do human-embryo and cloning research.

OCTOBER 2009
Hwang is found guilty of embezzlement and bioethics violations. Appeal continues.

2011
Canada grants Hwang a patent for the NT-1 cell line.

2012
Sooam scientists clone a coyote using a dog egg-cell donor and surrogate mother.

2013
Court tells the Korean Centers for Disease Control and Prevention to register the NT-1 cell line.

ongoing efforts to prove that the NT-1 cell line was in fact derived from an authentic clone could pave the way to future approvals.

Woo Suk Hwang has made some progress in convincing official bodies of NT-1's authenticity. In 2012, a Seoul court ordered the Korean Centers for Disease Control and Prevention to register the cell line — although this does not indicate its origins. The agency had initially refused on the grounds that eggs used in the experiments had been obtained unethically because donors were paid. But it was forced to relent because the eggs used to make NT-1 were obtained before the bioethics law banning the practice came into effect.

In 2011, Canada issued a patent to Sooam that refers to NT-1 as a cloned cell line. And Insung Hwang says that other patents are pending from some half a dozen of what he considers to be the "most symbolic" countries.

Getting recognition for NT-1 from the scientific community will be difficult, however. The paper in which NT-1 was reported[1] was clearly fraudulent and has been retracted. And SNU's finding that the line was a product of parthenogenesis has been backed up by an analysis[10] by George Daley, a stem-cell biologist at Harvard University in Boston, Massachusetts. He looked at thousands of DNA sites in the cell line and found that the chromosomes had recombination patterns strikingly similar to those of mouse parthenotes — evidence that Daley calls "unequivocal".

But a 2011 study[11] by Eui-Bae Jeung of Chungbuk National University in Cheongju, South Korea, argues that NT-1 does come from a true clone. This analysis is based on the similarities between the way the genes are methylated and expressed in the cell line and in cells from the nuclear donor.

Mahendra Rao, director of the US Center for Regenerative Medicine in Bethesda, Maryland, says that both analyses have their ambiguities. He says he believes that Daley's data are stronger, but that "more evaluation is required".

The most convincing evidence against NT-1 being a real clone might be that from Woo Suk Hwang's own lab. In 2003, when the researchers were preparing the paper, several tests indicated that NT-1 might be a parthenote, according to team leader Young-Joon Ryu. The Seoul prosecutor's report notes that another researcher, Sung Keun Kang, a former SNU professor and a right-hand man to Hwang, went back and altered the test results.

Many stem-cell scientists see Woo Suk Hwang's failure to publish NT-1 as a parthenote as a missed opportunity[12]. "He could have made a career studying parthenogenetic activation," says Schöler.

SECOND CHANCES

Among the public, opinions of Woo Suk Hwang are mixed. His actions have left many patients feeling betrayed — although some continued to support him with fervour. Susan Fajt, who was paralysed in a car accident and whom Hwang pledged to make walk again, continued to believe him after the fraud was revealed. "I talked with him for four hours. He had tears in his eyes. I don't think he would mislead anybody," she said in 2006. Fajt died in 2010.

But the scandal did not seem to have as disastrous an impact on support for stem-cell research worldwide as had been feared. Some in South Korea even credit the episode as partly responsible for a recent boom in stem-cell funding in the country (see *Nature* http://doi.org/qv5; 2012). "It was helpful," says Hyo-Soo Kim, a stem-cell scientist at SNU Hospital. "It drew attention and interest from government and ordinary people."

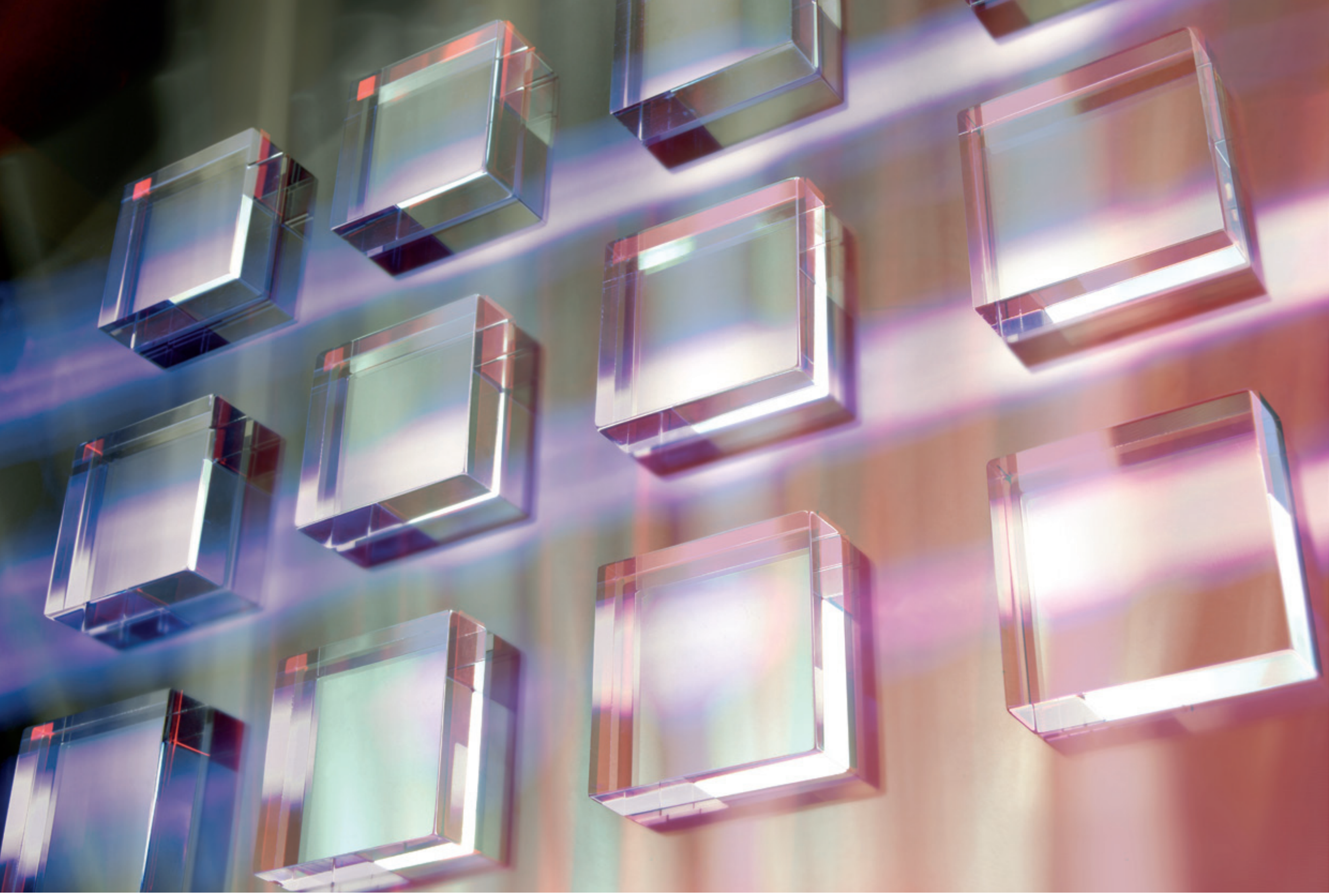
South Korea has now approved more stem-cell treatments than any other country. One such therapy, which uses stem cells derived from umbilical cords to tackle osteoarthritis, was approved in 2012 and is made by biotechnology firm Medipost in Seoul. Antonio Lee, chief executive of the company's US subsidiary, notes that immediately after the scandal the firm had trouble enrolling patients, "but at the same time it raised awareness among the general population about the potential of stem cells".

Overall, the case did not lead to a major erosion of public trust, says Bernd Pulverer, head of scientific publications at the European Molecular Biology Organization in Heidelberg, Germany, although it did raise important questions about how the problems went undetected for so long. "One clear issue that emerged was the danger of focusing such intense expectations to perform, and to funnel so much funding to one individual," he says. "At some point, it was clear that the stakes were simply too high for Dr Hwang to fail." For the research enterprise as a whole, he adds, "I am not sure anything changed fundamentally".

For Woo Suk Hwang, once at the centre of so much media attention, things have undoubtedly changed. In Sooam's chilly cafeteria he dines with a thick jacket on, chatting quietly to a handful of staff. He will greet a journalist and shake hands, but he does not want to talk about what happened. Hwang is increasingly surrounded by people who indulge him in that — offering a space for his ambitions to expand without constant reminders of his failures. When lunch is finished, he steals away, back to his ever-multiplying dogs and his hopes for redemption. ■

David Cyranoski *is a correspondent for* Nature *based in Shanghai, China.*

1. Hwang, W. S. *et al. Science* **303,** 1669–1674 (2004).
2. Hwang, W. S. *et al. Science* **308,** 1777–1783 (2005).
3. Lee, B. C. *et al. Nature* **436,** 641 (2005).
4. Jung, E.-M. *et al. Mol. Med. Rep.* **7,** 406–412 (2013).
5. Jeong, Y.-H. *et al. PLoS ONE* **8,** e63241 (2013).
6. Jung, E.-M. *et al. Mol. Med. Rep.* **6,** 239–245 (2012).
7. Jeong, Y. W. *et al. Int. J. Mol. Med.* **30,** 321–329 (2012).
8. Hwang, I. *et al. Reprod. Fertil. Dev.* **25,** 1142–1148 (2012).
9. Tachibana, M. *et al. Cell* **153,** 1228–1238 (2013).
10. Kim, K. *et al. Cell Stem Cell* **1,** 346–352 (2007).
11. Jung, E.-M. *et al. Int. J. Mol. Med.* **28,** 697–704 (2011).
12. De Sousa, P. A. & Wilmut, I. *Cell Stem Cell* **1,** 243–244 (2007).



ELEMENT SIX

The company Element Six produces pure diamonds with flaws at less than one part per billion.

FLAWED TO PERFECTION

Ultra-pure synthetic diamonds offer advances in fields from quantum computing to cancer diagnostics.

BY ELIZABETH GIBNEY

The 'magic Russian diamond', as some researchers have come to call it, was just 2 millimetres square, very clear and of a quality any jeweller would be happy to set in an expensive ring. Jörg Wrachtrup, a physicist at the University of Stuttgart in Germany, had spent much of 2005 looking for something just like it; his group finally found it by trawling through journals from the Russian Academy of Sciences, reading descriptions of the physical properties of such rare gems. But Wrachtrup wasn't interested in this diamond's beauty: what intrigued him was that the stone was very pure and perfectly flawed.

Inside the Russian gem, the regular diamond lattice of carbon atoms was interrupted on rare occasions by a nitrogen atom, with a neighbouring carbon atom also missing. Within each such hole, an extra electron could become trapped (see 'A useful hole'). Such impurities are not in themselves unusual. But Wrachtrup and others had theorized that, in some specific cases, electrons in these holes could prove the perfect medium for storing information for quantum computing — an effort to vastly speed up computing calculations by exploiting the fuzzy world of quantum mechanics. Unlike other candidates for such information storage, these defects in diamond should do their job at room temperature. To test the idea, Wrachtrup's lab split the diamond and sent half of it to Mikhail Lukin, a physicist at Harvard University in Cambridge, Massachusetts. By the end of 2006, both groups had shown that the Russian stone proved the theory correct[1,2]. "This diamond showed behaviour we had never seen before," says Wrachtrup.

Since then, the field has exploded. In 2005, just a handful of groups worldwide were working on the quantum possibilities of diamond; there are now about 75 in on the action. The Russian diamond has been cut up and divided between teams. Despite much searching, no other natural gems quite like it have been found. So researchers have turned their attention to making synthetic versions that are even better.

As more teams have entered into the game, so have ideas for potential applications. The same properties that make diamonds useful for storing quantum information also make them ideal for sensing magnetic fields with incredible precision, which could be used to eavesdrop on the processes in living cells in real time. Tiny sensors could provide cellular-level imaging with one billion billion times more sensitivity than

conventional magnetic resonance imaging (MRI), allowing investigators to map electrical activity in neurons or watch a cell's reaction to a drug.

"We're really solving problems we haven't been able to solve before," says Wrachtrup.

GROWN FROM SCRATCH

Diamond lovers are familiar with impurities for their ability to give the stones exotic hues: nitrogen can lend a yellow tone; boron turns them blue. What excites physicists is the 'spin' of the electrons trapped in such defects. That quantum property can be either up, down or somewhere in between — all at the same time. Such fuzziness is required for the basic unit of quantum computing, called quantum bits, or qubits. Unlike conventional computer bits that are either on or off, a qubit must have the capacity to exist in multiple states simultaneously, allowing a computer to perform parallel calculations.

Quantum properties such as spin are delicate, and are easily influenced by any outside interference. Diamond makes a great candidate for qubits because its rigid crystal structure helps to isolate and protect trapped electrons' fragile quantum states from random perturbations. The spin can, however, be manipulated by microwaves and read out using lasers.

Natural stones usually contain flaws at a level of about one-in-a-thousand atoms, which is much too many to make them useful for information storage: the defects are so close together that they interfere with each other, meaning that electrons cannot reliably hold any given spin state for long. By contrast, the Russian diamond contains fewer than one nitrogen atom per billion carbon atoms.

Back in 2005, Wrachtrup's tests showed that electrons in the Russian diamond could maintain a defined spin state for almost a millisecond; the only other set-ups able to maintain a spin state for this long were those that were super-cooled to near absolute zero and maintained under a high vacuum. Diamond allows scientists to change and read the quantum state of a single electron at room temperature using everyday lab equipment. "That was a bit of a game changer," says David Awschalom, a physicist at the University of Chicago in Illinois, who was one of the first investigators to work on quantum-grade diamonds.

Makers of synthetic quantum-grade diamonds try to achieve at least the same level of purity as that of the Russian diamond. Unlike stones made for jewellery or industrial cutting, these diamonds aren't grown by putting a lump of carbon under high temperature and pressure. Instead, gases such as methane and hydrogen are heated into a plasma, so that carbon atoms can be deposited onto a template layer by layer.

Some academic labs can make such diamonds themselves, but the major hub for research into this type of diamond production is the UK-based labs of the company Element Six, which has been synthesizing diamond — originally for cutting and drilling — for more than 50 years. Its business is booming. In July 2013, the company opened a £20-million (US$32.9-million) Global Innovation Centre in Harwell near Oxford, UK, to research and develop better diamond-production schemes for new applications. It now sells a few hundred off-the-shelf pure diamonds for quantum research each year, and its production of custom diamonds for specific projects has doubled annually since 2007, totalling 1,500 so far. In the lab where the custom diamonds are grown, a dozen machines hum away, teeming with feeding tubes that bring in the basic ingredients.

Element Six now sells an ultra-pure diamond with impurities lower than one part per billion, into which scientists can implant desired defects. Those cost about $1,000 each. For their custom diamonds, the company works with researchers to put flaws in precise layers and to control the levels of different carbon isotopes, which can also affect a diamond's properties. "Building it atom by atom gives you the ability to control impurities," says Geoff Scarsbrook, research and development operations manager at Element Six. The company provides these diamonds to researchers at no cost with the aim of developing intellectual property rights and opening promising new markets. "We are prepared to take quite a long view," says Scarsbrook.

That they have a long outlook is just as well. Producing a single qubit is one thing, but producing a functional quantum computer with many cooperating qubits is quite another — as researchers working with other materials have discovered. Since the mid-1990s, a few systems have emerged as the leading candidates for qubits, including ions trapped by an electromagnetic field and superconducting circuits, which must be super-cooled. Scientists working with these systems still struggle to deal with interference and to hook multiple qubits up into usable systems. So far, the world's best all-purpose quantum 'computers' are toy models comprising little more than a dozen qubits that can do small tasks such as factoring the number 15 (with one stand-out exception of a controversial, specialized type of quantum system, see *Nature* **498,** 286–288; 2013).

Diamonds show substantial potential, however, and some gems can now keep qubits protected from interference for long enough to do something useful, says Ronald Hanson, a nanoscientist at Delft University of Technology in the Netherlands. In 2012, for example, Lukin's team reported[3] achieving a lifetime for a diamond qubit of more than one second, on a par with what has been achieved in trapped atoms and about 10,000 times better than in superconducting circuits. To do this, his team used the trapped electrons' spin only as a messenger. To actually hold information they used the quantum-spin properties of the neighbouring impurities — such as a nitrogen atom or a carbon-13 isotope — which are about 1,000 times less sensitive to interference than the spins of electrons are. A trick to control the electrons' spin when they are not acting as a messenger can theoretically extend the qubit's lifetime by up to a minute.

QUANTUM BITS AND PIECES

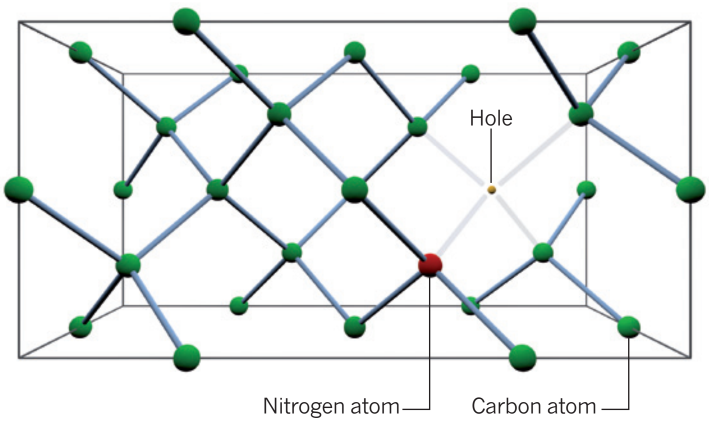
But connecting up the qubits — which involves 'entangling' their states so that they can work together to perform calculations — is a greater challenge. Wrachtrup's approach is to carefully position diamond defects about 20 nanometres apart in an array, so that the trapped electrons are close enough to entangle. Yet manufacturers have a hard time fabricating diamonds with such precisely placed defects. And the proximity of the imperfections means that the spin of each electron must be precisely controlled if the quantum states are to survive — something that is ever more difficult to achieve as the systems scale up in size.

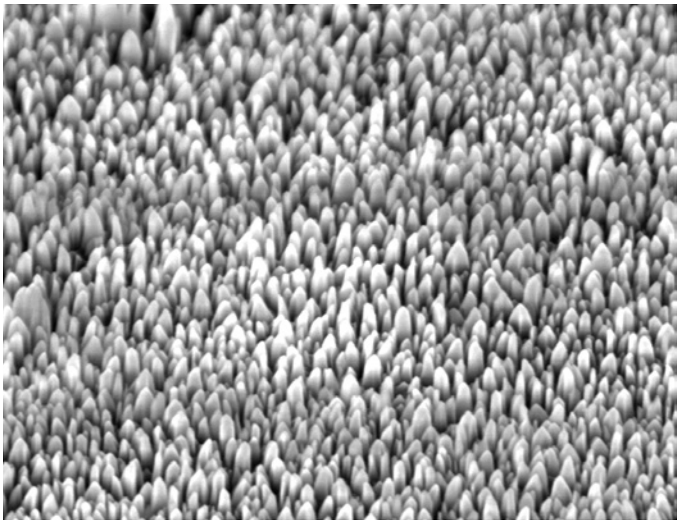
In an alternative approach, Hanson's team last year reported connecting up diamond qubits that are 3 metres apart by using flying intermediaries: photons that entangle with the electron spins and with each other[4]. That could prove particularly useful for a quantum network used to exchange information over long distances. But for Hanson's system to work, the qubits have to entangle in a much shorter time than the qubit lifetime. That means entanglement should happen many times a second; so far, Hanson and his collaborators have only succeeded in producing entanglement once every 10 minutes.

Physicists including Hanson, Lukin and Dirk Englund, an electronic

A USEFUL HOLE

The regular lattice of carbon atoms in a diamond crystal can be interrupted by a nitrogen atom sitting next to a missing carbon. Electrons trapped in the resulting hole can carry quantum information.





LUOZHOU L./ENGLUND GROUP/MIT

Nanometre-scale pillars etched from diamond can be used as magnetic probes.

engineer at the Massachusetts Institute of Technology in Cambridge, are trying to improve the entanglement rate by building tiny cavities and mirrors into thin films of diamond — this helps to bounce photons around and gives them more opportunities to interact with the electron qubits. Hanson thinks that this should make it possible to reduce entanglement times to fractions of a second. Teams are working on the best way to do this — some methods require thin films of diamond no more than a few hundred nanometres thick, laboriously ground from larger pieces. "It's very tedious to do," says Wrachtrup. "It's almost a work of art."

So far, the most sophisticated quantum-computing systems made of diamond — achieved separately by two different groups[5,6] — involve four entangled qubits. Scaling up to more than about ten will require a concerted engineering effort, says Wrachtrup. But diamond remains a viable option for quantum computing, with its greatest selling point being its ability to hold quantum information for long periods of time, at room temperature and without a vacuum. "It's really that combination that shows promise," says Hanson.

TINY MAGNETS

While researchers continue to battle with quantum computing, other applications for diamond could come to fruition more quickly. Some of the first researchers to explore the quantum properties of diamond realized that the way in which delicate spin states can be affected by their environment could be put to good use. The electrons' spins have a magnetic moment, which makes them act like tiny bar magnets that are sensitive to other nearby magnetic fields.

Sensing techniques such as MRI make use of a similar phenomenon — the spin inherent in hydrogen atoms — to spy inside the human body. But these require millions of atoms to get a signal. And for the greatest precision, the machines need to be cooled to very low temperatures. Diamond probes can be small enough and close enough to their target to pick up a signal from a single atom, at room temperature — the magnetic field of the atom affects the electrons' spin, which can be read with a laser.

Sensors that use large numbers of diamond defects to measure relatively large magnetic fields are already in development. At small scales there have been proof-of-principle studies, including work measuring spins in a drop of oil just 5 cubic nanometres in size[7] and even in a single molecule[8]. In 2011, a team led by Lloyd Hollenberg at the University of Melbourne in Australia put nanodiamonds into living cells, allowing scientists to study tiny magnetic changes within them[9]. Wrachtrup says that a diamond-based probe should eventually be able to image the structure of a complex molecule such as a protein, monitor activity in the brain or track the action of a drug in a cell — all without altering the living system being observed.

Lukin's group has also made use of nanodiamond probes to take temperature readings inside a cell to within a few hundredths of a degree[10], by monitoring the response of trapped electrons' sensitive spin to the expansion and contraction of a diamond lattice as it heats and cools. Nanodiamond probes should be able to detect changes of a few thousandths of a degree, and could be used to infer biological processes such as tumour metabolism.

However, making nanometre-sized ultra-pure diamonds for tiny probes is a real headache: the deposition method used by everyone, including Element Six, produces gems that cannot be separated from their template base. Most of the proof-of-principle work on nanodiamond probes has therefore been done using relatively impure diamonds, made through high pressure and temperature compression. This limits their sensitivity.

Englund's team has come up with a better means of production, which is now being commercialized by Diamond Nanotechnologies in Boston, Massachusetts, a company Englund set up with a former postdoc. They paint gold palladium dots over pure diamond and then etch away the bits of surface left exposed, producing a series of gold-topped diamond posts that his team calls 'nano-grass'. These can be mowed, and the gold tops easily removed, to produce individual, minuscule diamond pillars. When made in this way, the electrons trapped in the diamond defects hold their spin for 100 times longer than those in conventional nanodiamonds[11]. The firm is using these pillars to build a prototype magnetic field sensor that is sensitive enough to detect the field from just a few electrons.

DIAMOND DOUBLES

Researchers will need production improvements such as these if they want to squeeze all the promise out of diamonds. But there is still a long way to go to perfect precision doping of defects and the production of large thin films and complex diamond structures.

Fulfilling specifications like these is routine for many semiconductor materials, including silicon. So Awschalom's group is exploring whether it might be possible to reproduce the seemingly unique properties of diamond in such materials. In 2011, his group showed that silicon carbide — a relatively cheap semiconductor that has for decades been manufactured in large, thin films for use in electronics — can host defects in which bound electrons exhibit the same quantum quirks as in diamond[12]. His group has made silicon carbide qubits. But these lack the main advantage of diamond qubits: so far, the lifetime of trapped electron spin states in silicon carbide at room temperature is 20 times shorter than in diamond — too short for most practical applications.

Awschalom's group is among a number attempting various tricks to boost silicon carbide qubit lifetime, including purifying the isotopic composition of the material. And the team is collaborating with theorist and former colleague Chris Van de Walle at the University of California, Santa Barbara, to predict which defects in other crystalline materials — including gallium nitride, which is used in light-emitting diodes — might match diamond's properties. "It's definitely an extremely promising new direction," says Englund. "There could be many we just don't know about."

But for most researchers, diamonds remain the material of choice. With their extreme purity and controllable spin states, synthetic stones now outshine any natural gem. Even so, the original magic Russian diamond continues to prove its worth. "We still have pieces in the lab, and once in a while we use them," says Wrachtrup. "They're still among the best we have." ■

Elizabeth Gibney *is a reporter for* Nature *in London.*

1. Nizovtsev, A. P. *et al. Opt. Spectrosc.* **99,** 233–244 (2005).
2. Childress, L. *et al. Science* **314,** 281–285 (2006).
3. Maurer, P. C. *et al. Science* **336,** 1283–1286 (2012).
4. Bernien, H. *et al. Nature* **497,** 86–90 (2013).
5. Robledo, L. *et al. Nature* **477,** 574–578 (2011).
6. Dolde, F. et al. Preprint at http://arxiv.org/abs/1309.4430 (2013).
7. Staudacher, T. *et al. Science* **339,** 561–563 (2013).
8. Sushkov, A. O. *et al.* Preprint at http://arxiv.org/abs/1311.1801 (2013).
9. McGuinness, L. P. *et al. Nature Nanotechnol.* **6,** 358–363 (2011).
10. Kucsko, G. *et al. Nature* **500,** 54–58 (2013).
11. Trusheim, M. E. *et al. Nano Lett.* **14,** 32–36 (2014).
12. Koehl, W. F., Buckley, B. B., Heremans, F. J., Calusine, G. & Awschalom, D. D. *Nature* **479,** 84–87 (2011).

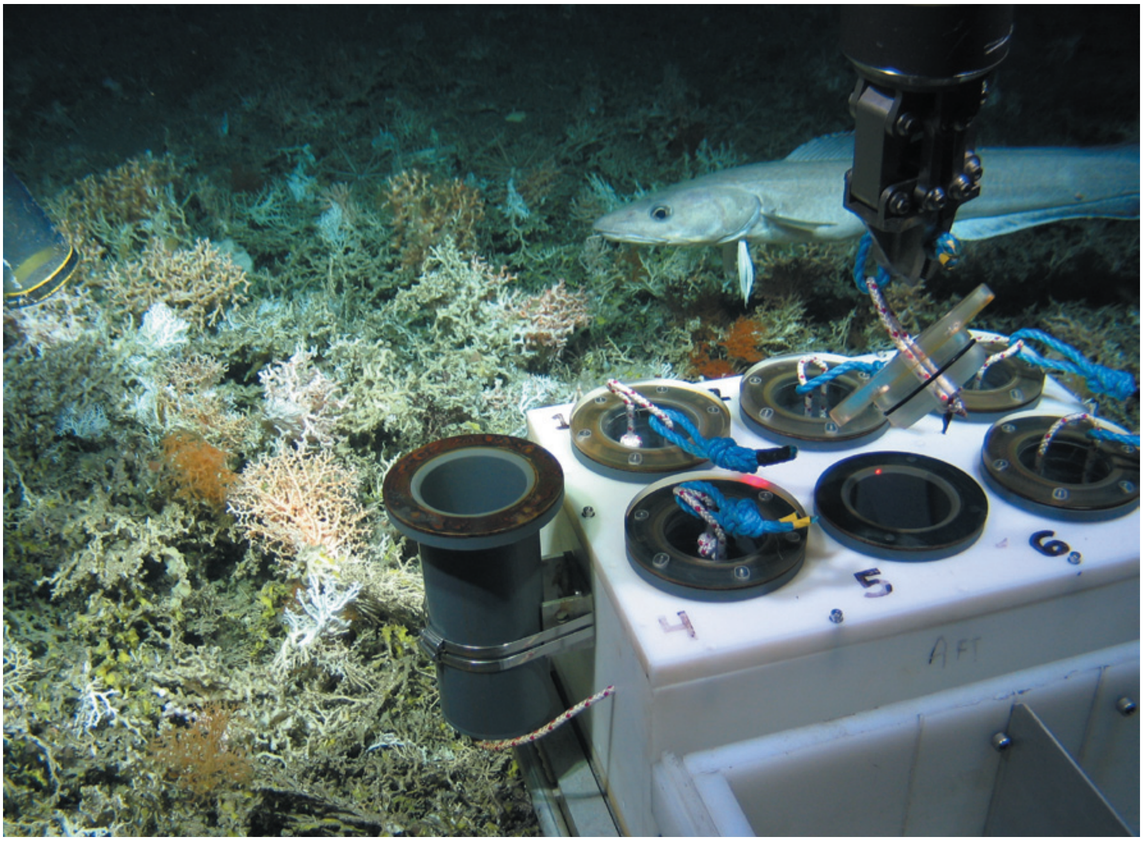
COMMENT

EDUCATION Raise scientists' social consciousness **p.477**

CONSERVATION Tim Flannery on Germaine Greer's foray into natural science **p.480**

PUBLISHING Don't shirk your peer-review duties, or science will suffer **p.483**

OBITUARY Janet Rowley, cancer-chromosomes pioneer, remembered **p.484**



HERIOT-WATT UNIV. CHANGING OCEANS EXPEDITION; CRUISE JC073, UKOA PROGRAMME, NERC

A remotely operated vehicle takes scientific samples on a coral complex in the northeastern Atlantic.

Protect the deep sea

Edward B. Barbier and colleagues call for governance and funds for deep-sea reserves and the restoration of ecosystems damaged by commercial interests.

More than 1 million square kilometres of the sea below 200 metres in depth are being ploughed by trawlers, according to estimates[1], and the next decade will see expansion of oil, gas and mineral extraction into deeper and deeper waters[2] (see go.nature.com/brhbll). At risk are ecosystems that contribute to the health and productivity of the ocean, that challenge our ideas of the extremes at which life can exist (such as hydrothermal vents), and that are habitats and nurseries for fisheries (seamounts, for example). Our knowledge of deep-ocean biodiversity only hints at thousands of undiscovered organisms and their benefits. Some threatened species, such as cold-water corals, have lifespans of hundreds or even thousands of years; habitats, including rock concretions called manganese nodule beds, can take millennia to form.

We call for formal governance structures and funds to be put in place by 2020 to create networks of deep-sea reserves that maintain and restore biodiversity and function in this vast and important biome[3]. To support these efforts, a global strategy must be framed under the aegis of national governments and an international body. For areas that are beyond national jurisdiction, the International Seabed Authority (ISA) is best suited to this task.

COSTS AND BENEFITS

Deep-sea restoration experiments have already begun. Cold-water corals from the northeastern Atlantic survive and grow in laboratories[4] and experimental reintroduction to the sea floor has proved successful, with 76% of corals surviving after three ▶

years[5]. Efforts are ongoing in the United Kingdom to develop 'coralbots', swarms of autonomous undersea vehicles to transplant and monitor coral fragments in the deep sea to overcome fishing damage.

But the potential effectiveness of large-scale restoration is unknown, and the precedents are not promising: after almost four decades of restoration, freshwater and coastal ecosystems still do not recover their full biodiversity and functionality. Repairing damage to and enhancing recovery of deep-sea ecosystems will be more expensive than for shallow ones by two to three orders of magnitude. For example, it could cost as much as US$75 million to restore one hectare of trawled seabed at the Darwin Mounds hummocks inhabited by corals at a depth of one kilometre in the Rockall Trough of the northeastern Atlantic[3].

"A key feature of a global strategy for protecting and restoring the deep sea should be the 'polluter pays' principle."

It is a price that many feel is worth paying. As well as oil, gas, mineral and biomedical resources, deep-sea ecosystems have other important functions, including roles in gas and climate regulation, and waste absorption and detoxification[6].

A 2007 study[7] revealed that the public in Ireland is willing to pay up to €10 ($14) per person to protect deep-sea corals from trawling so that the corals can provide raw materials for the biomedical industry, essential fish habitats and carbon sinks. Visitors and residents in the Azores, an Atlantic archipelago about 1,500 kilometres west of Portugal, expressed a willingness to pay €405–605 per person[8] to prevent 10–25% reductions in marine species richness in open waters, including the deep sea. In Scotland, survey[6] respondents were willing to pay £70 ($115) to £77 each to promote maximum deep-sea biodiversity conservation and develop new medicinal products from deep-sea species.

A GLOBAL STRATEGY

A key feature of a global strategy for protecting and restoring the deep sea should be the 'polluter pays' principle. That is, stakeholders who are most responsible for damages should fund deep-sea ecosystem reserves, research and restoration. These entities are likely to include mining, oil and gas, transportation and fishing companies.

However, implementation of this strategy will depend on whether the deep sea lies inside or outside national boundaries. For areas within national jurisdiction, the responsibility for restoration, protection and determining liability would fall on individual states. Governance in areas beyond national jurisdiction, where most of the deep sea lies, is currently divided according to sectorial activities — primarily fishing, shipping and mining. Because a universal authority to consider ecosystem protection, costs and benefits in international waters does not yet exist, adding a biodiversity-conservation agreement to the United Nations Convention on the Law of the Sea (UNCLOS) is under discussion, with a decision due in late 2015. Such a development is an essential first step for protecting the deep sea.

An important component of the 2015 UN General Assembly decision should be to either develop a new body to protect deep-sea biodiversity, or to extend the mandate of the ISA beyond mining to protect habitats from a wider range of regulated commercial industrial activities.

A key role of the Convention on Biological Diversity (CBD) is to provide scientific and technical advice to states and relevant authorities, so a close cooperation between the CBD and the ISA could be established even during such negotiations. This cooperation could apply the CBD's targets, which call for protecting and restoring 10% of the oceans, including the deep sea, by 2020.

RESTORATION FUND

To implement the updated UNCLOS agreement, a new fund of around $30 million per annum, perhaps managed by the ISA, is needed to cover conservation and restoration research, development and implementation for the deep sea in areas beyond national jurisdiction. This fund should start immediately after the 2015 decision, and comprise contributions from the national or private companies that undertake mining, transportation, fishing and other commercial activities that are harmful to sea-floor ecosystems. The ISA is charged with granting licences for deep-sea mining in the high seas and with sharing a proportion of the profits with the international community, primarily developing states. The fishing industry, by contrast, is accustomed to free access to deep-sea resources and is reluctant to pay for restoring seabed ecosystems affected by trawling[3].

A tax is an alternative to voluntary contributions. For example, the total catch value from high seas bottom trawling (HSBT) is $601 million per year for the 12 countries with major fleets[9]. A 1% tax on these revenues could raise $6 million annually (4% of the $152 million in subsidies that these countries currently give their HSBT fleets[9]). Current deep-water oil production (which is within national boundaries) is estimated at 5 million to 6.3 million barrels a day[10]. If states agreed, at current world prices of about $100 per barrel, a 1% royalty would generate between $5 million and $6.3 million a day.

Another alternative is an international finance facility, which would mobilize resources for deep-sea restoration from international capital markets by issuing long-term bonds to be repaid by donor countries over 20–30 years. For instance, the International Finance Facility for Immunisation (IFFIm) was launched in 2006 to provide funds for vaccinations, and it has so far received pledges of $6.3 billion for 23 years from nine donor countries (see www.iffim.org). A proposed Global Forest Finance Facility, based on the IFFIm, could serve as a model for a deep-sea finance facility.

SUSANNA STRÖMBERG/UNIV. GOTHENBURG

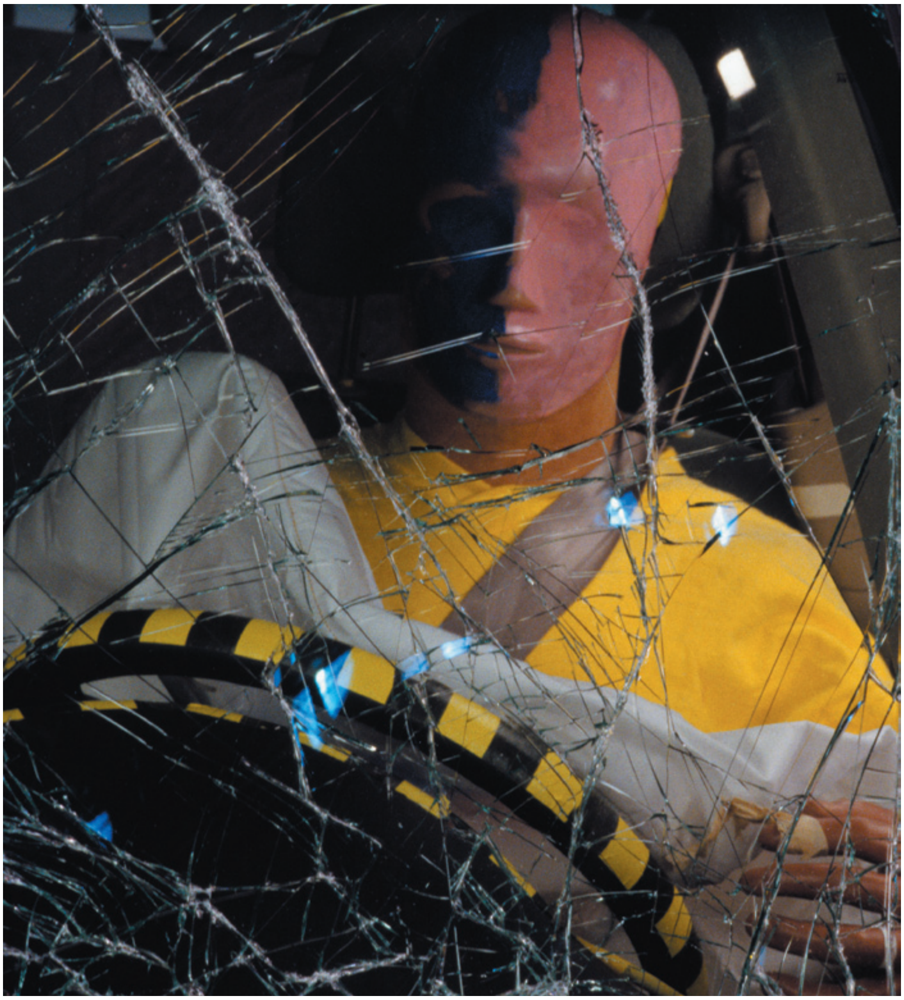
Coral (*Lophelia pertusa*) mounted on artificial reefs before use in restoration efforts near Sweden.

TIM WRIGHT/CORBIS

National governments, the international community and commercial interests should agree by 2015 on which mechanisms would work best to finance deep-sea protection and restoration, and by 2020, cooperate on implementing the fund. If we wish to continue to enjoy the benefits of deep-sea ecosystems, it is essential that we find ways to finance deep-sea research, reserves and restoration. ■

Edward B. Barbier *is professor of economics at the University of Wyoming, Laramie.* **David Moreno-Mateos** *is at the CNRS Centre of Evolutionary and Functional Ecology, Montpellier, France.* **Alex D. Rogers** *is in the Department of Zoology, University of Oxford, UK.* **James Aronson** *is at the CNRS Centre of Evolutionary and Functional Ecology, Montpellier, France, and at the Missouri Botanical Garden, St Louis, Missouri.* **Linwood Pendleton** *is senior fellow in the Ocean and Coastal Policy Program, Nicholas Institute for Environmental Policy Solutions, Duke University, Durham, North Carolina.* **Roberto Danovaro** *is in the Department of Life and Environmental Sciences, Polytechnic University of Marche, Ancona, Italy, and at the Anton Dohrn Zoological Station, Naples, Italy.* **Lea-Anne Henry** *is at the Centre for Marine Biodiversity and Biotechnology, School of Life Sciences, Heriot-Watt University, Edinburgh, UK.* **Telmo Morato** *is at the Institute of Marine Research in the Department of Oceanography and Fisheries, University of the Azores, Horta, Portugal, and at the Laboratory for Robotics and Systems in Engineering and Science, Portugal.* **Jeff Ardron** *is at the Institute for Advanced Sustainability Studies, Potsdam, Germany.* **Cindy L. Van Dover** *is in the Division of Marine Science and Conservation, Nicholas School of the Environment, Duke University, Beaufort, North Carolina.*
e-mail: ebarbier@uwyo.edu

1. Priede, I. G. *et al. ICES J. Mar. Sci.* **68,** 281–289 (2011).
2. Van Dover, C. L. *Nature* **470,** 31–33 (2011).
3. Van Dover, C. L. *et al. Mar. Policy* **44,** 98–106 (2013).
4. Strömberg, S. M., Lundälv, T. & Goreau, T. J. *J. Exp. Mar. Bio. Ecol.* **395,** 153–161 (2010).
5. Dahl, M. *Conservation genetics of* Lophelia pertusa. PhD Thesis, Paper V. Univ. Gothenburg (2013).
6. Jobstvogt, N., Hanley, N., Hynes, S., Kenter, J. & Witte, U. *Ecol. Econ.* **97,** 10–19 (2014).
7. Wattage, P. *et al. Fish. Res.* **107,** 59–67 (2011).
8. Ressurreição, A. *et al. Ecol. Econ.* **70,** 729–739 (2011).
9. Sumaila, U. R. *et al. Mar. Policy* **34,** 495–497 (2010).
10. Sandrea, R. & Sandrea, I. *Oil Gas J.* **108,** 48–53 (2010).



Early airbags were dangerous to women and children, having been designed for adult men.

Embed social awareness in science curricula

Separate ethics courses are not enough, argues **Erin A. Cech**. Understanding the public–welfare impacts of science and engineering is a core professional skill.

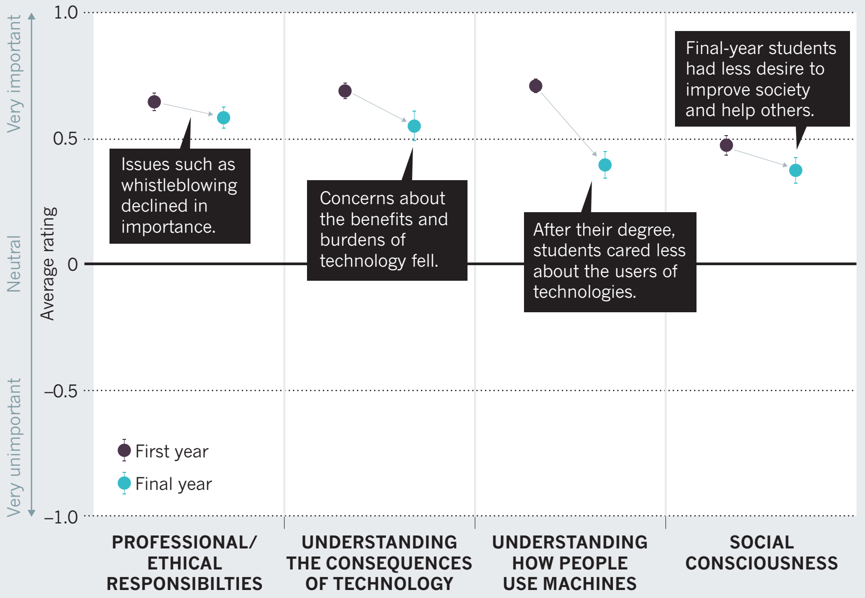
As a social scientist who is also trained as an engineer, I am puzzled by how often public-welfare and social-justice issues are viewed as irrelevant or tangential to 'real' technical work in science, technology, engineering and mathematics (STEM) professions. I carried out a study[1], the results of which suggest that university education exacerbates this culture of disengagement.

Between 2003 and 2008, I surveyed a total of more than 300 engineering students in four US universities — a large state college, an elite technical college, a small engineering-only university and a small private liberal-arts college. Following students from their first year to 18 months after their graduation, I found that, on average, they left their degrees less interested in public welfare than when they began.

The reverse should be true. STEM practitioners and educators increasingly recognize that those who understand the role of their profession in society are better at solving real-world problems[2]. Ethics courses for STEM students are proliferating. But adding a few courses is not enough. Social issues should be embedded throughout STEM curricula. Scientists and engineers must view the understanding of the social context ▶

SOCIAL DISENGAGEMENT

In a US survey, more than 300 engineering students rated four aspects of the social relevance of their work lower at the end of their undergraduate degrees than at the start.



of their work as a core professional skill.

This culture of disengagement is a concern because most STEM problems have cultural and political issues built into them[3–5]. The early design of safety airbags in cars, for example, was subject to gender bias. In 1993, the US National Highway Traffic Safety Administration dictated to manufacturers that the rate of force for airbag deployment had to be strong enough to protect an unbelted, average adult male. Car designers did not test their airbags on dummies of the average weight and stature of women or children; injuries and deaths followed[6].

A graduate student designing technology to read emotion in faces told me another story. On demonstrating the equipment to local school students, he realized that the method of recording changing expressions by reflecting light off faces did not work for people with dark skin. The technology had tested fine for everyone in the lab, but they were all light-skinned. "We didn't think to try it out on others who didn't look like us," he said.

The culture of disengagement also makes it more challenging to achieve equality within STEM. Discussions of power, exclusion, and inequality of women, lesbian, gay, bisexual, transgender and racial- or ethnic-minority individuals are typically seen as tangential — best left to diversity workshops and the like. But by standing aloof, we validate the existing power structures and unequal status quo.

My study[1] examined four attitudes among engineering students: the importance to them of their professional and ethical responsibilities (such as whistleblowing), of understanding the uneven consequences of technologies (such as nuclear technologies and the Internet), of understanding how people use machines, and of the desire to improve society and help others. Although most students rated these issues as 'important' rather than 'unimportant', they weighted them as more neutral in each subsequent year of their degrees (see 'Social disengagement').

"Public-welfare concerns should be incorporated into marked homework and exam problems."

The more-neutral scores lingered or worsened between graduating and entering the engineering workforce.

The findings suggest that this is not a simple tale of 'growing up' and losing naivety. It is clear that the curricular emphasis of engineering programmes had a significant effect on students' public-welfare beliefs. Students in programmes that played down the policy implications of engineering, for example, expressed less personal concern with professional and ethical responsibilities in the surveys.

PUBLIC WELFARE MATTERS

The diversity of educational approaches represented by these four universities suggests there is a broader problem across engineering education — and perhaps STEM in general. All four institutions require ethics courses and education in non-STEM subjects. Two of the colleges expressed commitments to producing 'well-rounded' engineers. It is not that these schools neglect engagement, but that wider culture instils in students the idea that social issues are not central to engineering.

I argue that the culture of disengagement in STEM is propped up by three ideological pillars. The first is depoliticization, the belief that science and engineering are 'pure' spaces free of political and cultural concerns[7]. Second is a technical–social duality, the assumption that technical knowledge and competencies have more value than social ones[8]. The third pillar is meritocracy, the belief that scientific professions are unbiased, with fair systems of advancement[7,9]. All three of these ideologies need to be challenged in the classroom and beyond.

What must be done? Public-welfare concerns should be incorporated into marked homework and exam problems. Rather than asking students to estimate the volume of an abstract pond, for instance, as one engineering programme does, students could work out the quantity of toxic materials produced by a plastics plant. This could open up discussions about possible effects on the community's water supply, about whether toxin levels were dangerous and, if so, how best to inform the community about potential dangers.

I believe that if even 10% of homework and exam questions required students to reflect on the social ramifications of research and results, scientists and engineers could reverse the slide into disengagement[10]. ■

Erin A. Cech *is assistant professor of sociology at Rice University, 6100 Main Street, Houston, Texas 77005-1892, USA.*
e-mail: ecech@rice.edu

1. Cech, E. A. *Sci. Technol. Hum. Val.* (in the press).
2. Schneider, J. *Eng. Stud.* **2,** 1–4 (2010).
3. Knorr Cetina, K. *Epistemic Cultures: How the Sciences Make Knowledge* (Harvard Univ. Press, 1999).
4. Barry, B. E. & Ohland, M. W. *Sci. Eng. Ethics* **18,** 369–392 (2012).
5. MacKenzie, D. *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance* (MIT Press, 1990).
6. Pacelli, L. *J. Contemp. Health Law Policy* **15,** 739–756 (1999).
7. Cech, E. A. In *Engineering Education for Social Justice: Critical Explorations and Opportunities* (ed. Lucena, J.) 67–84 (Springer, 2013).
8. Faulkner, W. *Soc. Stud. Sci.* **30,** 759–792 (2000).
9. Young, M. *The Rise of the Meritocracy* (Transaction, 1994).
10. Lucena, J. (ed.) *Engineering Education for Social Justice: Critical Explorations and Opportunities* (Springer, 2013).

CORRECTION

Owing to an editing error, the Comment piece by Amy W. Ando in 'The Endangered Species Act at 40' (*Nature* **504,** 369–370; 2013) wrongly stated that the ESA protected the American bison (*Bison bison*). The plains bison has never been listed under the act.

LYNDON MECHIELSEN/NEWSPIX/REX

Germaine Greer at Cave Creek, Australia, the rainforest she has worked to regrow on land partially cleared for farming.

CONSERVATION

Rewilding Oz

Tim Flannery celebrates Germaine Greer's foray into natural science — a chronicle of her rainforest-restoration project in a corner of Queensland.

Germaine Greer, feminist polemicist and fomenter of permanent revolution, has been hiding her light under a bushel. *White Beech* reveals her as one of the finest natural history writers to grace a page. This largely autobiographical work documents her restoration of around 60 hectares of montane (high-altitude) rainforest in southeast Queensland, Australia.

The project was balm for a deep hurt. She writes: "I had seen devastation, denuded hills, eroded slopes, weeds from all over the world, feral animals, open-cut mines as big as cities, salt rivers, salt earth, abandoned townships, whole beaches made of beer cans. Give me just a chance to clean something up, sort something out, make it right, I thought, and I will take it."

The property she settled on, at Cave Creek, had been logged, partially cleared for dairy cattle, then planted with bananas. But it retained extraordinary biodiversity. The white beech (*Gmelina leichhardtii*) of the title is one of Australia's most majestic rainforest trees, and Greer's labours commenced with the rescue of a dying forest giant that was slowly succumbing to overshading by lantana — plants of the verbena family, introduced from the Americas. Sunlight touched it, its first flush of leaves came in, and "the great old tree sent up a silent shout of victory and gushed torrents of blossom".

White Beech: The Rainforest Years
GERMAINE GREER
Bloomsbury Publishing: 2014.

Cave Creek hid other botanical treasures. Some of Australia's most endangered plants, including the smooth Davidson's plum (*Davidsonia johnsonii*) with its large plum-like fruit, and the corky-barked Glenugie karaka (*Corynocarpus rupestris*), thrive there. Greer sees it as her personal mission to protect such rarities.

White Beech meanders through history and botany like a vine looping through the canopy. From an exegesis on Agent Orange to explanations of botanical terms such as 'glaucous' (to describe a blue-grey waxy coating on leaves and stem), the book winds its way, guided by Greer's unique sense of where the light of truth lies. Botanists may bridle at its idiosyncrasies. "Nobody," Greer opines, "was less likely to give up the pernicious habit of calling plants after colleagues and friends than the egregious [Ferdinand] Mueller." By and large, Greer prefers descriptive names for plants. And yes, she peremptorily strips that distinguished pioneering botanist Ferdinand von Mueller of

his barony, refusing to acknowledge that his 'von' was both awarded and deserved.

Tracing the indigenous history of her patch of rainforest provides grist for Greer's extraordinary capacity for research. Having established that there were no Aboriginal owners (nobody went there because it was a 'story place', believed to be the haunt of vampire-like beings), Greer eventually, in 2011, gave the land to the UK charity Friends of Gondwana Rainforest. She explained, "If I have not learnt in my seventy-four years that to love and care for something you don't need to own it, then I have learnt nothing." In time it will be transferred to an Australian non-profit company.

White Beech is not without blemish. Referring to my area of expertise, mammalogy, I can report a handful of issues with her discussion of marsupials. The common planigale (*Planigale maculata*) is not in fact larger than the yellow-footed antechinus (*Antechinus flavipes*), and Greer seems to confuse the two species of quoll found in eastern Australia. I was, incidentally, also frustrated to learn that she gave short shrift to a researcher wishing to study the Hastings River mouse (*Pseudomys oralis*), one of Australia's rarest mammals. Other experts will doubtless find trifles to quibble with. But such peccadilloes are inevitable in a book that ranges so widely.

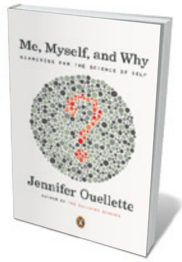
> ***"Give me just a chance to clean something up, sort something out, make it right, and I will take it."***

Many of the worst weeds at the site are already controlled or eradicated, and the rainforest is steadily taking over the pasture. Greer clearly has a vision of what a restored Cave Creek will look like; but nowhere does she spell it out in detail. Will every species that existed in the area in 1788 be returned? Will fire be used as a management tool? Such dilemmas dog all efforts at habitat restoration in Australia, because people — Aboriginals for 45,000 years, and Europeans for 225 — have hugely altered the land.

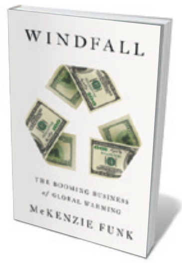
Greer put all she had into restoring Cave Creek. But can the immense effort of weeding alien species and afforestation be sustained through a small non-profit? Habitat restoration has become fashionable in Australia, and thousands now donate to organizations such as Bush Heritage and the Australian Wildlife conservancy, which restore habitat on a grand scale. It is hard to avoid the conclusion that some consolidation will be required if efforts like Greer's are to be sustainable. ■

Tim Flannery *is head of the Australian Climate Council. His latest book is* Among the Islands: Adventures in the Pacific.
e-mail: tim.flannery@textpublishing.com.au

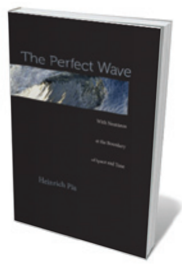
Books in brief



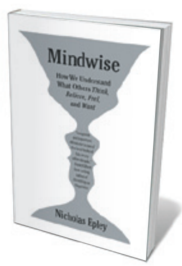
Me, Myself, and Why: Searching for the Science of Self
Jennifer Ouellette PENGUIN BOOKS *(2014)*
What sets us apart, genetically, neurologically and behaviourally? Science journalist Jennifer Ouellette's exploration of the "science of self" is an engrossing and often amusing tour of elite labs and edgy research. She is tested by US personal-genetics company 23andMe and in the belly of a magnetic resonance imaging machine in the lab of neuroscientist David Eagleman. She interviews behavioural psychologists, muses on digital doppelgängers, drops LSD and dips her toe into consciousness studies. Ultimately, she concludes, the self consists in what we make of our biological constraints.



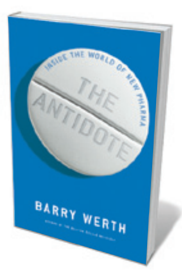
Windfall: The Booming Business of Global Warming
McKenzie Funk PENGUIN PRESS *(2014)*
This exposé of the powers and people that view global warming as an investment opportunity is darkly humorous and brilliantly researched. Journalist McKenzie Funk looks at the impacts deemed a windfall for "climate capitalists": melting ice, drought, sea-level rise and superstorms. He reports far and wide, on the oil-rich far north, where nations jostle as the ice retreats; blaze-prone California and its burgeoning band of firebreak specialists; water-rich South Sudan, where large tracts of foreign-owned farmland could become a gold mine as other regions dry up; and beyond.



The Perfect Wave: With Neutrinos at the Boundary of Space and Time
Heinrich Päs HARVARD UNIVERSITY PRESS *(2014)*
The ghostly neutrino — a mutable, almost massless particle that can pass through dense substances — stars in this scientific history. Theoretical physicist Heinrich Päs surfs the decades of dazzling research since Wolfgang Pauli first posited the particle in 1930. Päs revisits key theorists such as Ettore Majorana, and lays out the work of groundbreaking labs from Los Alamos in New Mexico, where Fred Reines and Clyde Cowan first detected neutrinos in the early 1950s, to today's vast IceCube neutrino observatory in Antarctica.



Mindwise: How We Understand What Others Think, Believe, Feel, and Want
Nicholas Epley KNOPF *(2014)*
Psychologist Nicholas Epley examines the "real sixth sense": inferring what others think, an ability essential in everything from high-level diplomacy to parenting. But as he shows, our conscious introspection is limited, and we tend to dehumanize others, as well as filter our perception of them through a screen of egotism. Epley sees the solution as the face-to-face work of open, honest communication — a tough call in a society addicted to texting and tweeting. Nuanced, authoritative and accessible.



The Antidote: Inside the World of New Pharma
Barry Werth SIMON AND SCHUSTER *(2014)*
In his follow-up to *The Billion-Dollar Molecule* (1994), Barry Werth re-enters the tough world of big pharma to trace the trajectory of drug company Vertex over the past two decades. The US-based company, once an upstart setting out to challenge the giants, now crafts promising treatments. Kalydeco (ivacaftor), for instance, treats cystic fibrosis by targeting the effects of a particular genetic mutation. A riveting mix of molecular science, big personalities — and big money. **Barbara Kiser**

MARTIN MEJIA/AP

Peruvian children with computers from the One Laptop Per Child Project.

DIGITAL DEVELOPMENT

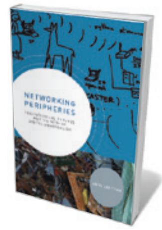
Wired cultures

John Gilbey discovers how Peru has leapfrogged standard models of technological roll-out to ignite social change.

Many see the development of Internet access in a country as a series of modest, incremental, linear changes in interactions with the technology. But in nations where people have leapfrogged landline Web access by adopting Internet-connected mobile devices, the changes can be transformative, both culturally and intellectually. Peru is a case in point. Among the fastest-growing economies in the world over the past decade, the country has seen rural mobile-phone use jump from 1.3% to 46.2% between 2004 and 2010. In 2011, three-quarters of Peruvian households had access to a mobile phone. Not all of this flux is technology-driven, but information technology and access to information are key factors in how the story is evolving.

In her unusual and fascinating *Networking Peripheries*, communications researcher Anita Say Chan uses two disparate but connected themes from Peru's recent history to shed light on the interplay between technology and social change at a development crossroads. One is the creation of a legal framework supporting open software. The other is the government's push to codify artisanal traditions — a process that has awakened Peruvians' interest in how intellectual property is managed in the Internet age.

As Chan reveals, the path that Peru has taken to develop information technology has significantly altered the aspirations of its rural population — perhaps even more than those of its urban elite — through an appreciation of "local realities". Focusing on the movement known as Free/Libre/Open Source Software (FLOSS), Chan shows us how, in 2005, Peru became one of the first countries to pass a law requiring public institutions to exercise technological neutrality — that is, to consider FLOSS options — when they decide which software to use. This was, Chan suggests, largely the result of populist campaigns by collectives of free-software advocates seeking to challenge the dominance of closed proprietary standards.



Networking Peripheries: Technological Futures and the Myth of Digital Universalism
ANITA SAY CHAN
MIT Press: 2013.

Chan also discusses the local rollout of the One Laptop Per Child project. This global initiative to provide every child with access to a low-cost, Internet-connected computer demonstrates the need for careful introduction of large-scale technology projects. In some instances, Chan shows, the delivery of technology seems to have been regarded by civic authorities as a complete outcome in itself, rather than a single step in a chain enabling development or reform through active engagement and learning. But as Chan explains, the efforts of teachers, trainers and fellow learners using specific local skills and contacts to carry the programme forward provide interesting, potentially replicable models for processes such as the development and sharing of locally relevant training materials.

These instances show Peruvians adapting new technologies in harmony with their own culture. By contrast, Chan reveals, government-backed artisanal projects may threaten local, traditional technologies with new but alien ones. The codification initiative aimed to standardize designs and techniques in ceramics production at 400 workshops in the town of Chulucanas, using larger-scale, non-traditional production methods. The result was a range of intellectual-property issues, such as the alleged exploitation of craftspeople and the loss of local cultural integrity through attempts at homogenization for a global market. The rise of a broad-ranging "information class" of growing ubiquity across the social landscape provides a real mechanism for resolving those issues.

Networking Peripheries reveals a Peru moving towards a dynamic and diverse future, in terms of both technology and culture. Chan clearly has a high regard, and legitimate concern, for the people and organizations she has engaged with — artisans, teachers, government officials, activists. Anyone who feels that they already understand the full impact of Internet technology on human culture and community may be surprised and intrigued by the first-hand material presented in this important text. ■

John Gilbey *teaches in the department of computer science at Aberystwyth University, UK.*
e-mail: gilbey@bcs.org.uk

Correspondence

Payback time for referee refusal

Through my current and past work as associate editor of several refereed journals, I have discovered a negative correlation between the number of papers that a scientist publishes per year and the number of times that that scientist is willing to accept manuscripts for review. In other words, the biggest consumers of peer review seem to contribute the least to the process.

There are two solutions to this situation. We could abolish peer review altogether, which would be tantamount to doing away with science as we know it. Alternatively, we can apply a well-known sociological principle, according to which no voluntary association can survive without incentives to increase compliance with its rules and penalties for disobedience. I therefore suggest that journals should ask senior authors to provide evidence of their contribution to peer review as a condition for considering their manuscripts. Such evidence should be easily verifiable in this age of data mining.

So, if you publish 10–20 research papers a year with the help of 30–60 referees, do your bit in return.

Dan Graur *University of Houston, Texas, USA.*
dgraur@uh.edu

Fraud is not the big problem

Whistle-blowers risk a huge personal backlash in exposing scientific misconduct (see, for example, D. Soeken *Nature* **505,** 26; 2014), but they can hope to correct only a tiny percentage of the published literature.

Since 1980, when MEDLINE started categorizing retractions, there have been 6,119 retracted papers, amounting to 0.03% of the 17.8 million published. Even if the majority of these retractions arise from misconduct (see, for example, F. C. Fang *et al. Proc. Natl Acad. Sci. USA* http://doi.org/jf5; 2012), this still affects only a very small proportion of the literature overall.

From alarming estimates derived from studies by Bayer (F. Prinz *et al. Nature Rev. Drug Discov.* **10,** 712; 2011) and Amgen (C. G. Begley and L. M. Ellis *Nature* **483,** 531–533; 2012) that some 60–70% of biomedical research papers may contain irreproducible results, it would seem that our time would be better spent investigating experimental irreproducibility rather than hunting down fraudsters.

William Gunn *Mendeley and the Reproducibility Initiative, California, USA.*
william.gunn@mendeley.com

Resolving soil pollution in China

On 30 December 2013, China's Ministry of Land and Resources reported that the country has 3.33 million hectares of farmland that are too contaminated to use. Given that this amounts to 2.5% of total arable land in China, more clarification is needed on the nature, extent, location and degree of soil contamination.

China is planning to invest billions of dollars in soil remediation in the coming years. But first the Chinese government should release detailed soil-pollution data so that the problem can be better understood and the sources of contamination brought under control by legislation. For example, regulations should be put in place with regard to the dumping of sewage and industrial wastewater in rivers or on cultivated land, and then strictly enforced.

The government also needs to make clear who is expected to supervise soil remediation and management, and when and how polluted land can be decontaminated under current laws.

Ruishan Chen *Hohai University, Nanjing, China.*
Chao Ye *Nanjing Normal University, China.*
yeover@163.com

Weighing up reuse of Soviet croplands

There is a pressing need to evaluate the trade-offs on abandoned Soviet croplands between food production, the provision of ecosystem services and biodiversity conservation (see *Nature* **504,** 342; 2013). Kazakhstan should be included in these assessments because, along with Russia, it commands some of the largest agricultural land reserves worldwide.

Trends in recultivation vary across the former Soviet Union. Reclamation in western Russia is only just starting, whereas Kazakhstan has reclaimed more than half of its abandoned cropland since 2000.

This intensified agricultural production is good for rural development and poverty alleviation (see M. Petrick *et al. World Dev.* **43,** 164–179; 2013). The implications for ecosystem services and biodiversity are less clear, however. Evidence is growing for biodiversity recovery (see, for example, J. Kamp *et al. Biol. Conserv.* **144,** 2607–2614; 2011) and for increased carbon sequestration on land depleted by intensive agriculture across the former Soviet Union.

Johannes Kamp *University of Münster, Germany.*
johannes.kamp@uni-muenster.de

Natural killers take on cancer

Your Outlook supplement on cancer immunotherapy (*Nature* **504,** S1–S17; 2013) focuses mainly on T cells as a promising immunotherapy tool. But natural killer (NK) cells, another type of immune cell, may also be suitable for treatment of some cancers and are currently being tested in clinical trials (see, for example, A. M. James *et al. Front. Immunol.* **4,** 481; 2013, and M. Cheng *et al. Cell. Mol. Immunol.* **10,** 230–252; 2013).

Unlike T cells, NK cells are not directly antigen-specific. However, they can use an antibody-dependent mechanism to kill tumour cells: the antigen-specific fragment of these antibodies recognizes molecules on the tumour-cell surface, which activates the NK-cell cytolytic machinery.

Another approach is to use antibodies that suppress the tumour-induced inhibition of NK cells. Also, anti-cancer agents such as lenalidomide act in part by modulating NK-cell function.

Jacques Zimmer *Public Research Centre for Health, Luxembourg.*
jacques.zimmer@crp-sante.lu

EDITOR'S NOTE

Nature has a strong history of supporting women in science and of reflecting the views of the community in our pages, including Correspondence. Our Correspondence pages do not reflect the views of the journal or its editors; they reflect the views only of the correspondents.

We do not endorse the views expressed in the Correspondence 'Publish on the basis of quality, not gender' (*Nature* **505,** 291; 2014) — or indeed any Correspondences unless we explicitly say so. On re-examining this letter and the process, we consider that it adds no value to the discussion and unnecessarily inflames it, that it did not receive adequate editorial attention, and that we should not have published it, for which we apologize.

Nature's own positive views and engagement in the issues concerning women in science are represented by our special from 2013: www.nature.com/women.

Philip Campbell, Editor-in-Chief, *Nature*

Janet Rowley

(1925–2013)

Geneticist who discovered that broken chromosomes cause cancer.

Janet Rowley, the 'matriarch of modern cancer genetics', transformed our understanding of cancer. In the 1960s she would cut out 'paper dolls' at the dining room table — but not for her children to play with. These photographs of human chromosomes eventually yielded discoveries that established the genetic basis of cancer and led to targeted cancer therapies.

Janet Davison was born in New York City on 5 April 1925 and spent most of her life in Chicago, Illinois. Encouraged by her mother, a high-school teacher and librarian, Janet received her bachelor's degree from the University of Chicago at the age of 19. Accepted subsequently to the university's medical school, she had to wait nine months to enrol because the school had already reached its quota of women: three in a class of 65. The day after her graduation in 1948, she married a classmate, Donald Rowley, who later became a distinguished pathologist.

In 1955, Janet began working part-time in a local clinic, where she treated children with Down's syndrome. The developmental disorder was linked in 1959 to an extra copy of chromosome 21, and Rowley became fascinated with inherited genetic diseases. When her husband took a sabbatical in England in 1961, she arranged to study with Laszlo Lajtha, a haematologist at the Churchill Hospital in Oxford. There, she began to examine chromosomes in the laboratory.

Returning to the United States in 1962, she secured a job with Leon Jacobson, a haematologist at the University of Chicago. Rowley asked for a darkroom, a microscope and a salary sufficient to pay a babysitter. Over the next decade, she scoured cells from people with leukaemia, looking for chromosomal abnormalities.

During a second sabbatical in Oxford, she perfected techniques to stain chromosomes, making it easier to identify them. She was the first to realize that bits of chromosomes in some human cancer cells had broken off and swapped places — a phenomenon known as translocation. The translocation that she identified between chromosomes 8 and 21 is now known to account for up to 12% of cases of acute myeloid leukaemia. She published the work in a singly authored paper in June 1973 (J. D. Rowley *Ann. Genet.* **16,** 109–112; 1973). The same month, she published a paper in *Nature* (J. D. Rowley *Nature* **243,** 290–293; 1973) that characterized a genetic abnormality found in people with chronic myeloid leukaemia.

UNIV. CHICAGO MEDICINE

Rowley showed that the 'Philadelphia chromosome', an aberrant version of chromosome 22 (named after the city where researchers identified the abnormality) was a genetic swap: the truncated chromosome 22 was accompanied by an elongated chromosome 9. Previously, a large genetic deletion had been thought to be involved. In 1977, she identified a third translocation, in people with acute promyelocytic leukaemia.

Her work was met with a chorus of scepticism and wonder that anyone would bother to study chromosome abnormalities, which were then considered to be an effect of disease rather than a cause. By the 1980s, however, each of the abnormalities had been molecularly characterized, revealing that translocations create 'fusion' proteins that drive cell growth. Since then, dozens of translocations have been found in other cancers.

For acute promyelocytic leukaemia, Rowley's discovery helped to uncover the mechanism behind an effective drug: retinoic acid. A derivative of vitamin A, the drug restores normal function to its disrupted protein receptor.

For chronic myeloid leukaemia, a disease that was once a death sentence, Rowley's discovery enabled work that led to new, effective treatments including imatinib, approved in the United States in 2001. She received numerous prestigious prizes for her research.

I first met Janet in 2000, when the efficacy of imatinib was well known. Aged 75, instead of dwelling on her seminal work on chronic myeloid leukaemia, she told me about new pathogenetic mechanisms of acute myeloid leukaemia that she was working on, and that she swam regularly in Lake Michigan and cycled to and from work.

Janet was also outspoken about her beliefs. Despite serving on former US President George W. Bush's Council on Bioethics, she was highly critical of the administration's policy that barred federal funding of embryonic stem-cell research. In 2009, she stood next to President Barack Obama when he lifted the ban.

In 2012, Janet, Nicholas Lydon and I were awarded the Japan Prize for work that led to imatinib. By the last evening of the week-long events surrounding the prize, I was spent. Janet, 30 years my senior and recovering from chemotherapy for ovarian cancer, was still going strong. I asked her how she managed. With a twinkle in her eye, she replied that I had had to chase my three children around while she had been able to rest.

That summed up Janet. She had incredible energy and curiosity and was gracious and humble, with the ability to make others around her feel good about themselves. ■

Brian J. Druker *is director of the Knight Cancer Institute at Oregon Health & Science University, Portland, Oregon, and a Howard Hughes Medical Institute investigator. He shared the 2012 Japan Prize with Janet Rowley and Nicholas Lydon.*
e-mail: drukerb@ohsu.edu

ECOLOGY

Good dirt with good friends

An analysis of data from forests across the planet reveals that the types of beneficial fungus with which tree roots associate determine the amount of carbon stored in soils. SEE LETTER P.543

MARK A. BRADFORD

In 1936, US President Franklin D. Roosevelt signed an act to conserve the "natural resources of the land", commenting[1]: "The history of every Nation is eventually written in the way in which it cares for its soil." Decaying organic matter improves soil health because it binds the soil, preventing erosion, and serves as a sponge that retains nutrients and water for plant growth. The amount of organic matter in soil is therefore an important determinant of its fertility and — because organic matter in the soil contains around three times as much carbon as the atmosphere[2] — of the magnitude of climate change[3].

Our understanding of what regulates the amount of soil organic matter is currently undergoing a conceptual upheaval[4]. On page 543 of this issue, Averill *et al.*[5] show that commonly assumed controls, such as temperature, do not explain why stores of organic matter differ across soils in temperate, tropical and boreal forests. Instead, the authors propose that a primary control is the types of fungus with which trees form mutually beneficial relationships.

Our health is increasingly seen to depend on the microorganisms that live in intimate association with us. It is the same for plants. Most species of land plant form relationships with soil microorganisms known as mycorrhizal fungi, which grow in and around their roots. The plants provide the fungi with simple carbon compounds (such as sugars) in exchange for nutrients such as nitrogen. The sugars fuel fungal activity, helping them to spread out from the plant into the soil, forming what is essentially an extended root network. The threads (known as hyphae) of this fungal web dramatically increase the surface area for nutrient uptake and exude enzymes to catalyse the decay of organic matter, releasing nutrients for plant growth.

However, as with our own friendships, mycorrhizal relationships vary in their costs and benefits to both partners. The relationship between arbuscular mycorrhizal (AM) fungi and trees, for example, could be likened to Facebook friendships in that both invest and receive little (they exchange small amounts of carbon and nitrogen). Ectomycorrhizal (EM) fungi and trees are more like best friends, however, demanding a lot but giving much in return. This key difference arises because EM fungi use the extra carbon they demand to access nitrogen from soil organic matter that is otherwise unavailable to plants, whereas AM fungi, like roots, primarily take up inorganic nitrogen from the soil (Fig. 1).

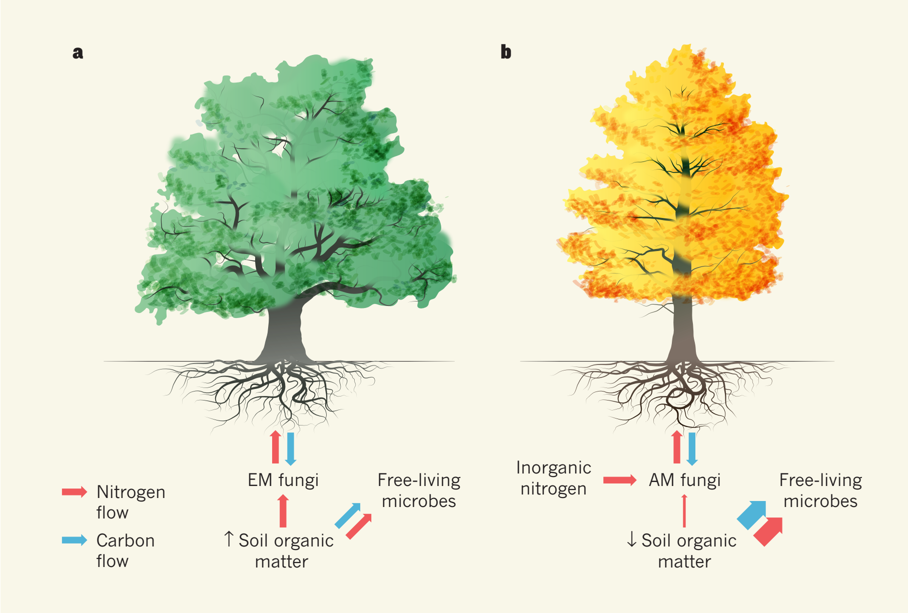


Figure 1 | Fungal types and soil organic matter. Some tree species, such as oaks, associate with ectomycorrhizal (EM) fungi (**a**), whereas others, such as maple, associate with arbuscular mycorrhizal (AM) fungi (**b**). In both relationships, the plants provide the fungi with carbon in exchange for nitrogen, but AM fungi primarily obtain inorganic soil nitrogen and EM fungi access nitrogen from organic matter. Averill *et al.*[5] show that the soils of forests dominated by EM associations have greater stores of organic matter than those dominated by AM associations. The authors propose that this difference arises because the use of organic nitrogen by EM fungi reduces the nitrogen available to free-living microbes that feed on organic matter, thereby slowing their activity and reducing the breakdown of organic matter.

Averill *et al.* studied the effects of these differing relationships on soil organic-matter stores by assembling a data set of the carbon and nitrogen content of soils around the world. The data revealed that ecosystems dominated by trees that form relationships with EM fungi store 1.7 times more carbon per unit of nitrogen than systems in which AM fungi dominate. The authors propose, in line with a previous hypothesis[6], that these richer carbon stores result from competition for nitrogen between EM fungi and free-living soil microorganisms that feed on organic matter — the EM fungi outcompete these microbes for access to nitrogen, retarding their activity and hence reducing losses of organic matter (Fig. 1).

The authors suggest that the consequences of these differences extend to the global scale. Climate warming is predicted to ramp up the metabolic activity of free-living soil microbes, increasing the carbon dioxide respired from soils to the atmosphere and thus creating a positive feedback loop that will amplify warming[3]. But the competition hypothesis suggests that this effect will be smaller in EM-dominated forests than in AM-dominated forests, because EM fungi will limit the organic-matter-degrading activity of free-living soil microbes.

Alternatively, the main reason for the higher organic matter in EM-dominated forests might be that trees in these systems allocate more carbon below ground to satisfy the greater demands of EM fungi[6]. This explanation is

consistent with the emerging idea that belowground plant inputs to soils are the dominant precursors for the formation of soil organic matter[4,7].

Pinpointing which mechanism explains Averill and colleagues' results will require more data and involve challenges common to all large observational data sets, including unobserved variables and spurious correlations. Perhaps different mycorrhizal associations reflect adaptations to environmental conditions, as opposed to being the cause of ecosystem differences. For example, in colder climates, where the cold slows the decay of organic matter and trees produce tough leaves that are hard to break down, EM fungi might dominate simply because of their ability to acquire nitrogen from organic matter[8,9].

In Averill and colleagues' analyses, the strength of the mycorrhizal effect depends on the amount of soil nitrogen. To investigate this dependency, I used their model results to calculate organic-matter stores in temperate and tropical forests, where the mycorrhizal types co-occur and where the authors conclude that EM-dominated forests have 1.3 times more carbon per unit nitrogen. At the low end of the authors' nitrogen-content range (0.2 kilograms of nitrogen per square metre), EM-dominated forests actually have less (0.96 times) carbon than AM forests, and at 1.0 kg N m^{-2}, below which many of the observations fall, they have 1.21 times more. It is not until soil nitrogen reaches values at the upper end of their observations (3 kg N m^{-2}) that carbon stores are 1.3 times greater in EM-dominated forests, a pattern consistent with the idea that the strength of the mycorrhizal effect is strongly dependent on soil-nutrient availability[6].

Despite the need to further explore such nuances, Averill and colleagues' findings have important implications for the way we manage land resources in the face of a changing carbon cycle and climate. We depend on model projections to inform strategies to preserve our natural resources, yet the relevant models have been developed on the basis of an understanding of soil dynamics that is increasingly shown to be wanting[10]. Climate, soil texture and plant productivity drive soil organic-matter storage in these models[11] but were found by Averill *et al.* not to play a determining part in organic-matter levels. Their finding that it is instead the relative dominance of trees associating with different mycorrhizal fungi that correlates with the amount of soil organic matter highlights the need to consider how local-scale biotic interactions shape global and regional-scale carbon dynamics. ■

Mark A. Bradford *is in the School of Forestry and Environmental Studies, Yale University, New Haven, Connecticut 06511, USA.*
e-mail: mark.bradford@yale.edu

1. Statement on Signing the Soil Conservation and Domestic Allotment Act, 1 March 1936. *American Presidency Project* http://www.presidency.ucsb.edu/ws/?pid=15254.
2. Jobbágy, E. G. & Jackson, R. B. *Ecol. Appl.* **10,** 423–436 (2000).
3. Denman, K. L. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. D. *et al.*) Ch. 7 (Cambridge Univ. Press, 2007).
4. Schmidt, M. W. I. *et al. Nature* **478,** 49–56 (2011).
5. Averill, C., Turner, B. L. & Finzi, A. C. *Nature* **505,** 543–545 (2014).
6. Orwin, K. H., Kirschbaum, M. U. F., St John, M. G. & Dickie, I. A. *Ecol. Lett.* **14,** 493–502 (2011).
7. Clemmensen, K. E. *et al. Science* **339,** 1615–1618 (2013).
8. Phillips, R. P., Brzostek, E. & Midgley, M. G. *New Phytol.* **199,** 41–51 (2013).
9. Johnson, N. C., Angelard, C., Sanders, I. R. & Kiers, E. T. *Ecol. Lett.* **16,** 140–153 (2013).
10. Wieder, W. R., Bonan, G. B. & Allison, S. D. *Nature Clim. Change* **3,** 909–912 (2013).
11. Bonan, G. B., Hartman, M. D., Parton, W. J. & Wieder, W. R. *Global Change Biol.* **19,** 957–974 (2013).

This article was published online on 8 January 2014.

SOLAR SYSTEM

Evaporating asteroid

The asteroid Ceres has been thought to contain abundant water. Observations acquired with the Herschel Space Observatory now show that this Solar System object is spewing water vapour from its surface. **See Letter p.525**

HUMBERTO CAMPINS & CHRISTINE M. COMFORT

Writing in this issue, Küppers *et al.*[1] report that Ceres — a dwarf planet or the largest asteroid in the Solar System, depending on the definition used — is releasing water vapour from its surface at a rate of about 2×10^{26} molecules, or 6 kilograms, per second. The presence and abundance of water in asteroids[2,3] are relevant to many areas of research on the Solar System, ranging from the origin of water and life on Earth to the large-scale migration of giant planets such as Jupiter.

Water has been suspected of being a significant component of Ceres for more than 30 years[4]. But it is only now that observations obtained by Küppers *et al.*, using the European Space Agency's Herschel Space Observatory, have allowed the direct identification of water molecules escaping from two regions on the surface of this object (Fig. 1). The authors' result backs up previous indirect observational evidence[5,6] for water in this planetary body, and is particularly timely given that NASA's Dawn spacecraft[7] will soon visit Ceres, fresh from its successful mission to another intriguing small world, the asteroid Vesta.

One of the most puzzling questions about the origin and evolution of asteroids is why Vesta and Ceres are so different. They are both located in the main asteroid belt, between the orbits of Mars and Jupiter, and their orbits are quite close to each other: about 2.4 and 2.8 astronomical units from the Sun, respectively (1 astronomical unit is the mean

Figure 1 | Artist's impression of the asteroid Ceres. Küppers *et al.*[1] have discovered water vapour emanating from two regions on the surface of the asteroid. (Figure adapted from an illustration by Chris Butler/SPL.)

Sun–Earth distance). Yet these objects are opposites in terms of their composition and appearance. Whereas Vesta has experienced extensive heating and volcanic eruptions that covered the entire asteroid, Ceres' surface and interior have not reached temperatures high enough to melt rocks.

Interestingly, a greater abundance of water in Ceres than Vesta may have been a crucial factor in producing the two bodies' radically different final states[8]. The source of the water vapour observed by Küppers and colleagues may be related to the process of heat dissipation that precluded the melting of rocks in Ceres. More specifically, one of the proposed production mechanisms for the water vapour being released from Ceres involves the melting of subsurface ice that then flows to the surface and evaporates into space. Water vapour has a high capacity to transport heat, and so, during the formation of Ceres about 4.6 billion years ago, the sublimation of water ice might have efficiently dissipated the interior heat into space. This would have stopped Ceres from ending up with an igneous surface like that of Vesta.

If this is indeed what happened during the formation of the two asteroids, one may ask: why did Ceres form with (and why does it still contain) more water than Vesta? It is most likely that Ceres formed in a colder outer region of the nascent Solar System than Vesta, beyond the snow line — the distance from the young Sun at which temperatures were low enough for water to form ice. But this hypothesis raises the question of why Ceres and Vesta are so close to each other now. It has been suggested that, soon after the formation of the asteroids and the planets, mixing of material from the inner and outer regions of the Solar System occurred. Such mixing would have been caused by migration of the orbits of Jupiter and the other giant planets[9], and that could have moved Ceres and Vesta from distant formation sites to their current locations.

One of the first clues that giant planets in the Solar System could undergo significant migration came from the discovery[10] in 1995 that certain giant exoplanets are closer to their hosts than Mercury is to the Sun — orbiting at distances at which they could not have formed. The best explanation for these 'hot Jupiters' is that they formed far from their host star and that later their orbits reduced dramatically.

Planetary migration has since been used to explain several puzzling observations. For example, the migration of Jupiter may have been responsible for the different compositional groups observed within the asteroid belt[9] and for a period of extensive impacts — known as the Late Heavy Bombardment — that occurred about 4 billion years ago[11–13]. According to this scenario, as the giant planets migrated, they disturbed populations of small rocky and icy bodies (asteroids and comets), which hit the early Earth and Moon. These small bodies delivered organic molecules and water to Earth. Hence, early impacts by asteroids and comets might have played a considerable part in the origin and evolution of life on our planet.

Küppers and colleagues' detection of water vapour around Ceres and, more generally, our knowledge of Ceres and Vesta, are consistent with emerging views of how giant-planet migration and other related processes shaped the Solar System's early history. But the pieces of the puzzle of Solar System formation do not fit perfectly, and more is likely to be discovered through further studies of the miniature worlds that we call asteroids. ■

Humberto Campins *and* **Christine M. Comfort** *are in the Department of Physics and Astronomy, University of Central Florida, Orlando, Florida 32816-2385, USA.*
e-mail: campins@physics.ucf.edu

1. Küppers, M. *et al. Nature* **505,** 525–527 (2014).
2. Campins, H. *et al. Nature* **464,** 1320–1321 (2010).
3. Rivkin, A. S. & Emery, J. P. *Nature* **464,** 1322–1323 (2010).
4. Lebofsky, L. A. *Mon. Not. R. Astron. Soc.* **182,** 17–21 (1978).
5. A'Hearn, M. F. & Feldman, P. D. *Icarus* **98,** 54–60 (1992).
6. Rivkin, A. S., Howell, E. S., Vilas, F. & Lebofsky, L. A. in *Asteroids III* (eds Bottke, W. F. Jr, Cellino, A., Paolicchi, P. & Binzel, R. P.) 235–253 (Univ. Arizona Press, 2002).
7. Russell, C. T. & Raymond, C. A. *Space Sci. Rev.* **163,** 3–23 (2011).
8. McCord, T. B. & Sotin, C. *J. Geophys. Res.* **110,** E05009 (2005).
9. Walsh, K., Morbidelli, A., Raymond, S., O'Brien, D. & Mandell, A. *Meteorit. Planet. Sci.* **47,** 1941–1947 (2012).
10. Mayor, M. & Queloz, D. *Nature* **378,** 355–359 (1995).
11. Tsiganis, K., Gomes, R., Morbidelli, A. & Levison, H. F. *Nature* **435,** 459–461 (2005).
12. Morbidelli, A., Levison, H. F., Tsiganis, K. & Gomes, R. *Nature* **435,** 462–465 (2005).
13. Gomes, R., Levison, H. F., Tsiganis, K. & Morbidelli, A. *Nature* **435,** 466–469 (2005).

STEM CELLS

Sex specificity in the blood

Haematopoietic stem cells, from which blood cells originate, are shown to respond to oestrogen and divide more frequently in female mice than in males, probably preparing females for the increased demand for blood in pregnancy. SEE LETTER P.555

DENA S. LEEMAN & ANNE BRUNET

Males and females exhibit differences not only in reproductive organs, but also in sexually dimorphic tissues such as the mammary gland, brain and muscle. In such tissues, the activity of stem cells, which self-renew and produce differentiated cells for tissue maintenance and repair, differs between males and females[1–4]. A fundamental yet unexplored question is whether the stem cells of tissues without conspicuous sex differences, such as the blood or gut, also exhibit sexually dimorphic function. On page 555 of this issue, Nakada *et al.*[5] find that haematopoietic stem cells (HSCs), which form the blood and immune system, do differ between male and female mice. The authors show that female HSCs respond to long-range oestrogen signals in a manner that seems to help mothers meet the haematopoietic demands of pregnancy.

HSCs reside in the bone marrow and produce all blood cells, which in turn mediate processes ranging from immunity to clotting to oxygen transport. Nakada and colleagues find that, under basal conditions, female HSCs and their immediate progeny, multipotent progenitor cells (MPPs), divide more frequently than male HSCs, and generate more erythroid progenitors (the cells that give rise to red blood cells).

Despite the increased frequency of division in female HSCs, males and females have the same basal number of HSCs and a similar cellular composition in the bone marrow and spleen (an organ colonized by haematopoietic cells). The authors suggest that female HSCs undergo more asymmetric divisions in which one daughter cell remains a stem cell and the other differentiates along the red blood cell lineage, and that these newly produced erythroid progenitors undergo cell death at a higher frequency (Fig. 1). These differences may explain why the sexual dimorphism of HSCs has not previously been observed.

During pregnancy, however, the authors observed a further increase in HSC proliferation, an expansion of the number of HSCs in the bone marrow and spleen, and more erythroid cells in the spleen. Thus, it seems that female HSCs may be 'primed' for the increased demand for blood during pregnancy.

Nakada *et al.* identify oestrogen as the causal agent for HSC sexual dimorphism (Fig. 1). They find that ovariectomy or pharmacological inhibition of aromatase (an enzyme necessary for oestrogen synthesis) reduced the percentage of proliferating HSCs and MPPs in females, whereas injection of oestradiol (the

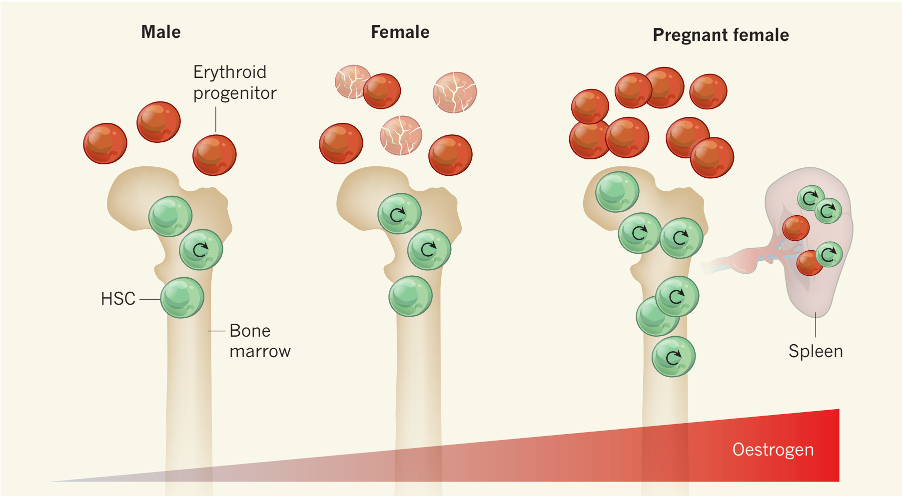


Figure 1 | Oestrogen regulates stem-cell cycling. Nakada *et al.*[5] report that haematopoietic stem cells (HSCs) respond to oestrogen signals, causing female HSCs to divide (circular arrows) more frequently than male HSCs, and produce more erythroid progenitors, which give rise to red blood cells. Despite this difference, the composition of the male and female bone marrow and spleen remains similar under basal conditions, because in females a higher percentage of erythroid progenitors undergo cell death. During pregnancy, however, oestrogen levels increase, leading to an expansion of the HSC population in the bone marrow and spleen, and more erythroid cells in the spleen. This response seems to play a key part in meeting the haematopoietic demands of pregnant females.

predominant oestrogen in females) increased HSC proliferation and drove erythropoiesis in normal and ovariectomized females, as well as in males. Although oestradiol injection was recently reported to induce cell proliferation in a subset of bone-marrow cells[6], Nakada and colleagues' study extends this observation by showing that physiological levels of oestrogen are sufficient to specifically influence HSCs. The authors also show that *Esr1*, the gene encoding oestrogen receptor-α (ERα), is highly expressed in HSCs and is necessary for the enhanced proliferation of female HSCs under steady-state conditions and during pregnancy. Furthermore, they find that oestradiol increases proliferation of wild-type HSCs, but not of *Esr1*-deleted HSCs, strongly suggesting that HSCs respond to oestrogen through ERα.

These findings raise the exciting possibility that the sensing of sex hormones by organs that are not sexually dimorphic may be necessary to orchestrate biological functions such as pregnancy. A key remaining question is whether this oestrogen-induced haematopoietic expansion is necessary for successful pregnancy or for maternal or fetal health. The mechanisms of action and target genes of ERα in HSCs are also not known, and their elucidation will contribute to our understanding of oestrogen-induced HSC proliferation and how it compares with oestrogen-induced responses in other stem cells.

Many genetically modified and naturally occurring mouse strains that have increased HSC proliferation exhibit premature exhaustion of HSC pools[7], so it will be interesting to investigate whether females show more HSC depletion than do males over long periods of time. HSCs are relatively quiescent cells, and this state is thought to protect them from the damage caused by cellular respiration and DNA-replication errors. However, it has also been suggested that DNA repair is more effective in cycling HSCs than in quiescent ones[8,9]. It would be worth testing whether HSCs exhibit sex-specific protection or repair mechanisms that allow female HSCs to sustain increased proliferation. Such exploration could reveal mechanisms by which HSCs might sustain increased proliferation without premature exhaustion or transformation.

Several studies have begun to address the questions of when and how tissue-specific stem cells are mobilized and coordinated by long-range signals in response to the body's systemic needs. Stem-cell function is affected by systemic signals, including those resulting from diet, circadian rhythm, exercise, mating and pregnancy[2]. During pregnancy, for example, increases in oestrogen and progesterone levels coordinate an expansion of mammary stem cells, which is required for remodelling of the mammary gland[3]. And increases in the hormone prolactin stimulate the production of pancreatic β-cells[10] and the proliferation of neural stem cells[4], which may have roles in responding to the increased metabolic load of the pregnant female and in maternal recognition of offspring, respectively. Nakada *et al.* have now introduced the concept that long-range signals act not only in response to specific systemic needs, but also under basal conditions to keep stem cells in a primed state, ready to act when pregnancy is initiated.

Sexual dimorphism in stem cells is understudied, and many stem-cell studies have been performed on only one sex or analysed without distinguishing between sexes. Nakada and

NATURE

50 Years Ago

Fifty Years of X-ray Diffraction (edited by P. P. Ewald) — The book … provides an extremely refreshing commentary on varieties of organization of scientific research, seen through the eyes of the very young … Wyart coming to Mauguin's laboratory, where there were only two elderly professors, with two elderly servants who kept the place clean and were rather worried about the mess he made, preparing crystals. There is young Schubnikov, trying to buy a lathe to cut crystal sections in Sverdlovsk in 1920 — and, since its price doubled in a few days, spending a million roubles of his own money to get it. There is young Mosley, who could not stop an experiment once he had started it and knew where to get a meal in Manchester at 3 o'clock in the morning … Wars and revolutions necessarily enter into the memories recorded here, though there are only casual references to the adventurous lives led by many crystallographers — to Carl Hermann, for example, working out the crystal structures of the urea adducts in prison … Perhaps most moving is the account of the reunion that took place after the Second World War when Laue himself came to London and met, after a long separation, crystallographers from all over the world. One has, throughout these pages, a very strong impression of being among a very united group of friends — united, as Bijvoet says, by a delight in crystals. **Dorothy Hodgkin**
From *Nature* 25 January 1964

100 Years Ago

The late Capt. Scott's original journals written during his expedition to the south pole, have been placed on view in the manuscript department of the British Museum.
From *Nature* 22 January 1914

colleagues' work suggests that attention to sex differences will be essential for understanding basic stem-cell biology, diseases associated with stem cells, and regenerative medicine. It is tempting to speculate that the authors' findings could provide explanations for sex differences that are poorly understood. For example, the enhanced engraftment of human HSC transplants in female compared to male mice[11] could be attributable to oestrogen. In addition, the slower erosion of telomeres (specific sequences at the ends of chromosomes that shorten with each cell division) observed in human bone-marrow transplants from female donors[12] might reflect a greater requirement of female HSCs to guard against exhaustion.

Further studies of sexually dimorphic regulation of stem cells could also provide insight into why some human diseases have a sex bias. Myelodysplastic syndromes and cytopenias, for example, which are characterized by reduced production of mature blood cells, often erythroid cells[13], have a higher incidence in men. This bias is interesting in light of Nakada and colleagues' identification of lower basal levels of HSC proliferation and erythropoiesis in male mice. Thus, their study imparts the unexpected lesson that consideration of sex in the context of disease pathogenesis and therapeutics may prove valuable even in diseases of seemingly non-sexually dimorphic organs. ■

Dena S. Leeman *and* **Anne Brunet** *are in the Department of Genetics, the Cancer Biology Program, and the Glenn Laboratories for the Biology of Aging, Stanford University, Stanford, California 94305, USA.*
e-mail: abrunet1@stanford.edu

1. Nakada, D., Levi, B. P. & Morrison, S. J. *Neuron* **70,** 703–718 (2011).
2. Shingo, T. *et al. Science* **299,** 117–120 (2003).
3. Ray, R. *et al. Mol. Med.* **14,** 493–501 (2008).
4. Deasy, B. M. *et al. J. Cell Biol.* **177,** 73–86 (2007).
5. Nakada, D. *et al. Nature* **505,** 555–558 (2014).
6. Illing, A. *et al. Haematologica* **97,** 1131–1135 (2012).
7. Orford, K. W. & Scadden, D. T. *Nature Rev. Genet.* **9,** 115–128 (2008).
8. Mandal, P. K., Blanpain, C. & Rossi, D. J. *Nature Rev. Mol. Cell Biol.* **12,** 198–202 (2011).
9. Mohrin, M. *et al. Cell Stem Cell* **7,** 174–185 (2010).
10. Nielsen, J. H., Svensson, C., Galsgaard, E. D., Møldrup, A. & Billestrup, N. *J. Mol. Med.* **77,** 62–66 (1999).
11. Notta, F., Doulatov, S. & Dick, J. E. *Blood* **115,** 3704–3707 (2010).
12. Baerlocher, G. M. *et al. Blood* **114,** 219–222 (2009).
13. Jain, A. & Naniwadekar, M. *BMC Hematol.* **13,** 10 (2013).

ELECTRONICS

Protecting the weak from the strong

A thin engineered surface has been developed that can protect sensitive electronic systems from strong signal interference, allowing them to communicate effectively with external antennas.

GEORGE V. ELEFTHERIADES

Conventional microwave absorbers have several uses, including in controlling unwanted signal reflections from antennas and radars, and in protecting sensitive communications electronics from strong interfering signals. Typically, these absorbers are made from a host material loaded with conducting iron- or carbon-based particles to absorb microwave energy and convert it into heat. A common feature of these absorbers is that the percentage of microwave power that is absorbed does not depend on the power of the incoming waves. Therefore, such devices cannot differentiate between strong signals, which can be harmful to communications electronics, and the weak signals that are needed for wireless communications with antennas. Writing in *Physical Review Letters*, Wakatsuchi *et al.*[1] report an engineered surface (metasurface) that can absorb high-power microwave pulses but allow weaker signals to propagate.

What enables Wakatsuchi and colleagues' metasurface to distinguish low- from high-power signals is the use of semiconductor diodes in the structure. A diode is a basic electronic building block that conducts high-power signals but cannot respond to low-power ones. Therefore, a diode can be thought of as a switch that is 'on' for strong signals but remains 'off' for weak signals. Such behaviour is nonlinear because the output power is not proportional to the input power. This is what enabled the authors to make such a power-dependent microwave absorber. Specifically, the team used a periodic array of copper patches printed on a thin dielectric (insulating) substrate backed by a copper ground plate. The array's unit cell consists of a single patch and six electronic components — four diodes, a resistor and a capacitor — which are soldered between patches (Fig. 1a, b). The overall thickness of the metasurface absorber is just 1.52 millimetres, whereas the spatial period of the array is 18 mm.

Now, the reader may wonder why the authors used four diodes, and what the role is of the capacitor and resistor in each unit cell. Well, the answer is that the authors' metasurface absorber is more subtle and interesting than my description so far. A single diode

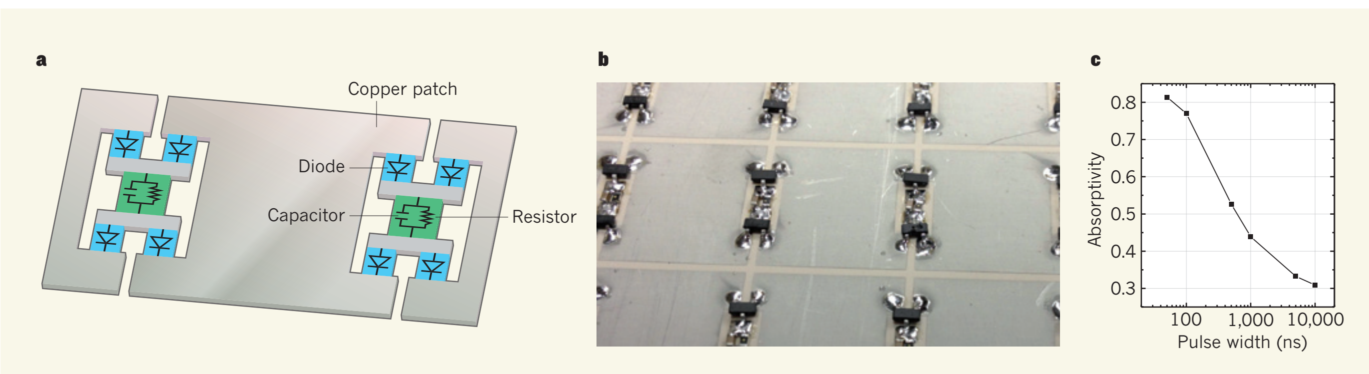


Figure 1 | A metasurface microwave absorber. a, Structure of the periodic metasurface designed by Wakatsuchi and colleagues[1]. Each unit cell comprises a copper patch, four diodes, a resistor and a capacitor. **b**, Photo of the metasurface, illustrating its periodicity. Each patch size is 17 mm × 17 mm, and there is a 1-mm gap between patches. **c**, The absorptivity of the metasurface depends on the width of the incoming pulse: short pulses are highly absorbed, whereas long pulses are poorly absorbed. (**a** and **c** adapted from ref. 1.)

has asymmetric conductivity even when it is switched on: it conducts well in one direction but poorly in the reverse. Ideally, for an absorber application — in which as much microwave energy should be converted into heat as possible — the absorber should conduct well in both directions when the power level is above a certain threshold. The use of four diodes does exactly this. It is a common topology used in electronics to make devices known as full-wave rectifiers. These have universal use; for example, they are common in power supplies that convert a.c. (alternating current) power to d.c. (direct constant current) to charge laptops and everyday consumer electronics. With signals being conducted in both directions, the metasurface absorber becomes more efficient and converts most of the incident energy into heat. And this last statement explains the presence of the resistor in the unit cell: the energy is eventually absorbed by the resistor, which naturally converts electrical energy into heat. But why involve a capacitor in the unit cell?

A capacitor is a simple device (two parallel conducting plates with a dielectric in between) that stores electrical energy. The rate at which this storage takes place is controlled by a quantity known as the *RC* time constant, which has units of time. This quantity endows the metasurface absorber with one of its most subtle and striking features: pulse-shape-dependent absorption. Energy is stored in the capacitor when a pulse impinges on the absorber. This energy is then dissipated in the resistor (the capacitor is discharged through the resistor) in the time between two successive pulses. In this way, the amount of absorption depends not only on the incoming power level of the pulse, but also on its shape. This is the reason that the title of Wakatsuchi and colleagues' paper is "Waveform-dependent absorbing metasurface". Shorter pulses lead to high absorption, whereas longer pulses are not absorbed well (Fig. 1c).

To put the results in context, in the past three years there has been a renewed interest in the field of metasurfaces. However, most of the metasurfaces described so far have linear behaviour (the input and output signals are proportional to each other) and do not contain electronic devices such as diodes. For example, a metasurface has been designed[2] to refract light by controlling the phase shift (delay) that the light undergoes as it propagates. Another example is a surface[3] with a tailored absorption achieved using antennas made of metal nanoparticles, the absorptivity of which does not depend on the incident power. Moreover, a passive metasurface — one that consumes but does not produce energy — has been engineered[4] to make thin cloaks for small dielectric cylinders, and a more general thin active cloak was reported last year[5]. However, Wakatsuchi and colleagues' metasurface is unique because its absorption performance is nonlinearly dependent on the shape and power level of the incoming wave.

The authors' waveform-dependent absorber has applications in several disciplines. For example, it could be applied to the skin of military vehicles or aircraft to protect sensitive electronics from strong electromagnetic-pulse threats, while allowing the electronics to communicate with external antennas. Moreover, one could imagine applying these absorbers to protect computer-network electronics such as those in data centres from strong interfering signals while allowing the electronics to operate properly. And one may foresee sensor or signalling applications based on the recognition of the width and power level of incoming pulses. In the authors' words, these metasurface absorbers could potentially create "new kinds of microwave technologies and applications". ■

George V. Eleftheriades *is in the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario M5S 3G4, Canada.*
e-mail: gelefth@ece.utoronto.ca

1. Wakatsuchi, H., Kim, S., Rushton, J. J. & Sievenpiper, D. F. *Phys. Rev. Lett.* **111,** 245501 (2013).
2. Yu, N. *et al. Science* **334,** 333–337 (2011).
3. Moreau, A. *et al. Nature* **492,** 86–89 (2012).
4. Rainwater, D. *et al. New J. Phys.* **14,** 013054 (2012).
5. Selvanayagam, M. & Eleftheriades, G. V. *Phys. Rev. X* **3,** 041011 (2013).

This article was published online on 15 January 2014.

CLIMATE SCIENCE

A resolution of the Antarctic paradox

A combination of observational data and modelling reveals the potential significance of the north and tropical Atlantic Ocean in driving change in Antarctic winds and sea ice on decadal timescales and longer. **SEE LETTER P.538**

JOHN KING

In recent years, the polar regions have provided some striking examples of rapid environmental change. Perhaps the most notable of these has been a reduction of more than 30% in the summer extent of Arctic Ocean sea ice since the late 1970s[1]. In the Antarctic, the pattern of change has been more complex. Although the extent of Antarctic sea ice has fallen significantly in some regions, it has increased in others, leading to a slight rise in overall winter-ice extent[2] (Fig. 1). Establishing the drivers of these changes has proved challenging, and the observed increase in Antarctic sea-ice extent — seemingly paradoxical in a warming climate — has frequently been used to question the widely accepted view that recent climate change is primarily anthropogenic in origin. A study by Li and colleagues[3], reported on page 538 of this issue, suggests that long-term warming of the north and tropical Atlantic may be the ultimate cause of the observed changes in the Antarctic. The authors' findings imply that growing Antarctic sea ice may be consistent with a generally warming Earth.

In contrast to ice in the Arctic Ocean, which is confined by the surrounding continents, Antarctic sea ice is largely free to drift with the wind and ocean currents. Its extent is therefore strongly influenced by the pattern of surface winds around the continent. Much of the year-to-year variability in Antarctic sea ice is captured by a pattern known as the Antarctic dipole[4], which is characterized by anomalies in ice extent of opposing signs in the Bellingshausen Sea and the western Ross Sea. The ice anomalies are a result of wind variations associated with changes in atmospheric-pressure patterns around the Antarctic. It is well established that these changes are connected to anomalies in sea surface temperature (SST) in the tropical Pacific Ocean[4,5] through the generation of large-scale atmospheric waves by deep convection in the tropical atmosphere. These waves, known as Rossby waves, can propagate to polar latitudes and influence the atmospheric circulation there. Much of the year-to-year variability in Antarctic sea ice can thus be attributed to variability in tropical Pacific SSTs.

The observed long-term trends in ice extent — retreat in the Bellingshausen Sea and compensating advance in the western Ross Sea — strongly resemble the Antarctic-dipole pattern and closely match long-term trends in winds over the Southern Ocean[6]. It would therefore be natural to look first to the Pacific as the driver of this change. However, long-term trends in tropical Pacific SSTs are small and cannot explain the observed trends in the Antarctic. In their study, Li and co-authors highlight instead the potential importance of the Atlantic in driving change in the Antarctic. This result is motivated by observations[7] showing

that, in contrast to the tropical Pacific, north Atlantic SSTs have warmed significantly since 1979. The authors demonstrate that warmer Atlantic SSTs drive anomalous Southern Ocean winds that are consistent with the observed regional trends in Antarctic ice extent. Although forcing from the tropical Pacific dominates the variability of Antarctic winds and sea ice on interannual timescales, Atlantic forcing becomes important on decadal and longer timescales, in which Pacific SST variability is smaller.

By establishing a chain of attribution linking warming of the tropical and north Atlantic with trends in Antarctic atmospheric circulation and sea ice, Li and colleagues' work helps to resolve the paradox of growing Antarctic sea-ice extent over a period when global mean temperature has increased. The researchers have also demonstrated that global climate models can simulate the connection between Atlantic SSTs and Antarctic winds. Why, then, have climate models such as those used in last year's fifth assessment report by the Intergovernmental Panel on Climate Change been unable to reproduce the observed regional pattern of change in Antarctic sea ice[8]?

Two reasons suggest themselves. First, the recent warming of the Atlantic is the result of a combination of anthropogenic forcing and natural internal variability of the climate system[7]. Only the effects of the former can be predicted in a deterministic way by climate models, with natural variability appearing as 'noise' in the climate-model simulations. Second, sea ice is one of the most challenging elements of the Earth system to model. The rate at which it forms or melts is controlled by the small difference between large fluxes of heat from the atmosphere and the ocean, and its distribution is strongly influenced by winds and ocean currents. Small biases in the models' representation of the atmosphere or ocean can thus translate into large errors in modelled sea ice.

Although accurate modelling of Antarctic sea-ice trends will require a realistic representation of the processes connecting Atlantic SSTs and Antarctic winds, this might not be sufficient. Given the importance of Antarctic sea ice to the Southern Ocean marine ecosystem, and its role in driving global ocean circulation by the production of ocean bottom water, understanding its behaviour and improving its representation in climate models must remain a high priority for climate scientists. ■

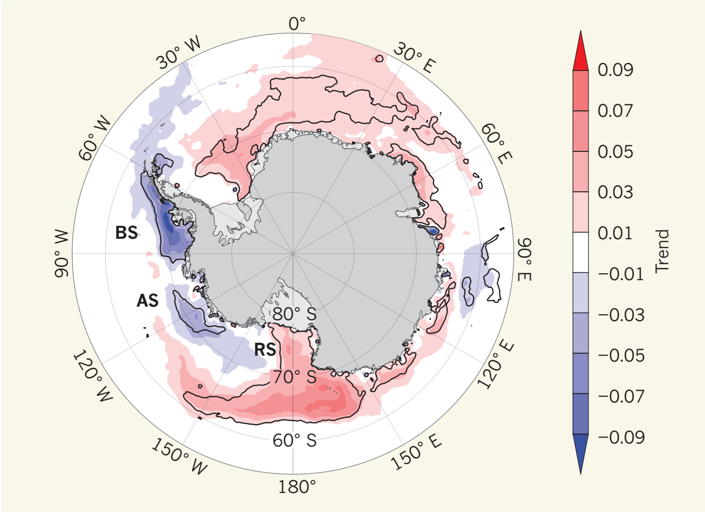


Figure 1 | Trend in Antarctic sea-ice coverage. The trend is expressed as the change in fractional ice coverage per decade and is calculated for the period 1979–2012. The bold lines enclose areas where the change is statistically significant at the 5% level. Ice cover has declined in the Bellingshausen Sea (BS), but compensating increases in the western Ross Sea (RS) have led to a slight overall increase in cover. These trends in ice cover are consistent with changes in winds driven by a deepening of the climatological low-pressure centre over the Amundsen Sea (AS). Li *et al.*[3] suggest that the changes in ice cover and winds have been caused by increased temperatures in the tropical and north Atlantic. (Data: National Snow and Ice Data Center, Boulder, Colorado. Image: British Antarctic Survey.)

John King *is at the British Antarctic Survey, High Cross, Cambridge CB3 0ET, UK.*
e-mail: jcki@bas.ac.uk

1. Stroeve, J. C. *et al. Clim. Change* **110,** 1005–1027 (2012).
2. Turner, J. *et al. Geophys. Res. Lett.* **36,** L08502 (2009).
3. Li, X., Holland, D. M., Gerber, E. P. & Yoo, C. *Nature* **505,** 538–542 (2014).
4. Yuan, X. & Martinson, D. G. *Geophys. Res. Lett.* **28,** 3609–3612 (2001).
5. Turner, J. *Int. J. Climatol.* **24,** 1–31 (2004).
6. Holland, P. R. & Kwok, R. *Nature Geosci.* **5,** 872–875 (2012).
7. Ting, M., Kushnir, Y., Seager, R. & Li, C. *J. Clim.* **22,** 1469–1481 (2009).
8. Turner, J., Bracegirdle, T. J., Phillips, T., Marshall, G. J. & Hosking, J. S. *J. Clim.* **26,** 1473–1484 (2013).

HIV

Not-so-innocent bystanders

The discovery that most CD4⁺ T cells killed during HIV infection die through a process known as pyroptosis may provide long-sought explanations for HIV-associated T-cell depletion and inflammation. SEE ARTICLE P.509

ANDREA L. COX & ROBERT F. SILICIANO

The first paper to describe AIDS reported that patients had very few $CD4^+$ T cells in their blood[1]. Depletion of this crucial subset of immune cells is now known to be a key feature of the disease, but the mechanisms responsible for their loss have remained unclear. Particularly mysterious has been the observation that HIV-1 infection results not only in the death of activated, productively infected $CD4^+$ T cells (those in which the virus successfully replicates) but also in 'bystander' $CD4^+$ T cells that do not seem to be infected. On page 509 of this issue, Doitsh *et al.*[2] show that most $CD4^+$ T cells depleted during HIV-1 infection are abortively infected cells that die through pyroptosis — a cell-death mechanism that is distinct from apoptosis and necroptosis[3].

HIV-1 replication in productively infected $CD4^+$ T cells kills them quickly, within one to two days[4,5]. This direct killing is apparent during acute infection, when virus levels are high and massive depletion of $CD4^+$ T cells occurs in the gastrointestinal tract[6]. However, in the absence of treatment, most of the $CD4^+$ T-cell loss associated with the infection occurs during the prolonged asymptomatic phase between the acute stage and the development of AIDS. During this period, the number of activated, productively infected $CD4^+$ T cells is low, suggesting that the infection may promote death of quiescent (non-activated) cells.

Levels of immune activation are high in untreated HIV-1 infection, perhaps reflecting the translocation of microbial products across a compromised gastrointestinal barrier[7], and it is commonly assumed that this immune activation is responsible for $CD4^+$ T-cell loss. Perhaps the best evidence for this comes from studies of simian immunodeficiency virus infections, in which there is high virus replication, but little immune activation or $CD4^+$

T-cell depletion[8]. Nevertheless, the mechanistic link between immune activation and CD4+ T-cell depletion has remained unclear.

Doitsh and colleagues suggest that this link may lie in the manner of cell death. Using cultures of human cells isolated from the spleen or tonsils, they demonstrate that more than 95% of CD4+ T cells that die following HIV-1 infection are quiescent cells that undergo pyroptosis. Only a small proportion of the dying cells were activated, productively infected CD4+ T cells undergoing apoptosis (Fig. 1). Apoptosis depends on the activation of the cell-signalling molecule caspase-3, whereas pyroptosis is triggered by inflammasome-activated caspase-1. Inflammasomes are multiprotein cytoplasmic complexes that integrate pathogen-triggered signalling pathways and then recruit and activate inflammatory caspase molecules. Pyroptosis results in lysis of the cell and release of the cytoplasmic contents into the extracellular space, and is highly inflammatory.

Productive HIV-1 infection involves the virus binding to the T-cell surface and entering the cell. There, the viral RNA is reverse transcribed to DNA and integrated into the host-cell genome, resulting in replication of the virus. If this process is aborted before integration and viral replication occur, the infection is termed non-productive. Doitsh and colleagues previously demonstrated[9] that there is selective depletion of CD4+ T cells in which incomplete viral DNA transcripts accumulate following abortive infection. The same research group also recently identified interferon-γ-inducible protein 16 (IFI16) as the host-cell DNA sensor that triggers this cell death[10].

To verify that most CD4+ T-cell depletion occurring during HIV-1 infection is mediated by pyroptosis, the authors treated cells with inhibitors of caspase-3 or caspase-6 (important in apoptosis), or of receptor-interacting protein kinase enzymes (important in necroptosis), and found that these treatments did not prevent most of the CD4+ T-cell loss. Also consistent with pyroptosis and the associated release of intracellular contents into the extracellular milieu was the presence of the cytoplasmic enzyme lactate dehydrogenase in the cell-culture supernatants. *In vivo* evidence for pyroptosis came from the detection of caspase-1 in quiescent CD4+ T cells in the paracortical zone that surrounds the region of activated CD4+ T and B cells in HIV-1-infected lymph-node tissues. The authors did not detect caspase-1 in the zone of activated CD4+ T cells or in uninfected tissue.

Caspase-1 activation is known[11,12] to induce secretion of the highly inflammatory cytokine proteins interleukin-1β (IL-1β) and IL-18, which contribute to inflammatory conditions such as atherosclerosis and metabolic syndromes[11–14]. Doitsh and colleagues show that IL-1β release also occurs after infection with HIV-1, and that this requires caspase-1 activation (Fig. 1). Finally, the authors show that pyroptosis induced by HIV-1 can be prevented with VX-765, a caspase-1 inhibitor that has previously been tested in people with chronic epilepsy and psoriasis, and found to be safe and well tolerated. VX-765 treatment inhibited caspase-1 activation, IL-1β secretion and CD4+ T-cell death in HIV-1-infected cell cultures.

These findings raise the possibility of reducing immune activation and inflammation in response to chronic viral infections through caspase-1 inhibition. The research also suggests two new approaches to improve HIV-1 therapy: the use of antiretroviral agents that act early in the viral life cycle to block abortive infection, and the use of agents that inhibit caspase-1. Combination therapy with multiple classes of antiretroviral drugs is the standard of care for patients infected with HIV-1, and this therapy effectively suppresses viral replication. Suppression of caspase-1 activation may not be necessary if combination therapy prevents abortive infection as well.

Although Doitsh *et al.* do not report IL-18 levels in their study, this cytokine is generally produced along with IL-1β after inflammasome activation, and elevated levels are associated with inflammatory conditions[13–15]. Serum IL-18 levels, which are known to be high in HIV-1 infection, are reduced by antiretroviral therapy[14,15]. Thus, it remains to be seen whether caspase-1 inhibitors will add to existing antiretroviral therapy for the treatment of HIV-1 infection. Either way, the implication of pyroptosis in CD4+ T-cell depletion is a new explanation for this 30-year-old mystery in HIV-1 pathogenesis. ■

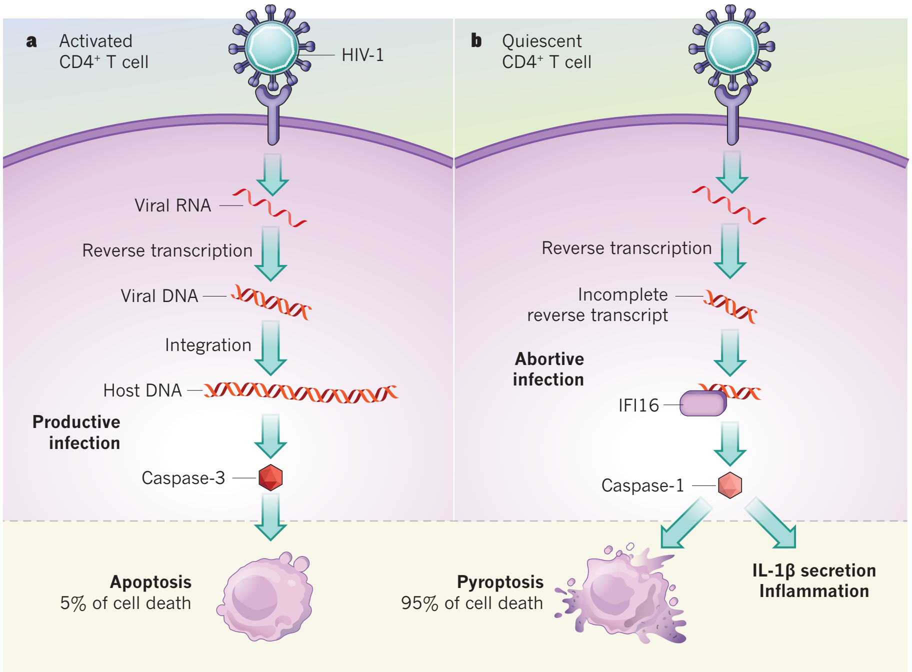


Figure 1 | CD4+ T-cell death during HIV-1 infection. **a**, Productive infection of a CD4+ T cell with HIV-1 involves viral entry to the cell, reverse transcription of viral RNA to DNA and integration of viral DNA into the host-cell genome. Following one or two days of viral replication, the activated, infected cell dies through apoptosis, mediated by the action of the enzyme caspase-3. Only about 5% of the CD4+ T cells that die after HIV-1 infection are activated, productively infected cells. **b**, Doitsh *et al.*[2] show that most CD4+ T-cell deaths result from caspase-1-mediated pyroptosis in non-activated (quiescent) CD4+ T cells that have undergone abortive infection, during which incomplete viral DNA transcripts remain in the cells. These transcripts are sensed by the cellular DNA sensor IFI16, which leads to caspase-1 activation, resulting in the secretion of the highly inflammatory cell-signalling molecule IL-1β and pyroptosis.

Andrea L. Cox *and* **Robert F. Siliciano** *are in the Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21218, USA.* **R.F.S.** *is also at the Howard Hughes Medical Institute, Baltimore.*
e-mails: acox@jhmi.edu; rsiliciano@jhmi.edu

1. Gottlieb, M. S. *et al. N. Engl. J. Med.* **305,** 1425–1431 (1981).
2. Doitsh, G. *et al. Nature* **505,** 509–514 (2014).
3. Miao, E. A. *et al. Immunol. Rev.* **243,** 206–214 (2011).
4. Ho, D. D. *et al. Nature* **373,** 123–126 (1995).
5. Wei, X. *et al. Nature* **373,** 117–122 (1995).
6. Veazey, R. S. *et al. Science* **280,** 427–431 (1998).
7. Brenchley, J. M. *et al. Nature Med.* **12,** 1365–1371 (2006).
8. Silvestri, G. *et al. Immunity* **8,** 441–452 (2003).
9. Doitsh, G. *et al. Cell* **143,** 789–801 (2010).
10. Monroe, K. M. *et al. Science* http://dx.doi.org/10.1126/science.1243640 (2013).
11. Latz, E. *et al. Nature Rev. Immunol.* **13,** 397–411 (2013).
12. Lamkanfi, M. *et al. J. Leuk. Biol.* **82,** 220–225 (2007).
13. Mallat, Z. *et al. Circulation* **104,** 1598–1603 (2001).
14. Iannello, A. *et al. Curr. HIV Res.* **8,** 147–164 (2010).
15. Watanabe, D. *et al. Viral Immunol.* **23,** 619–625 (2010).

ARTICLE

doi:10.1038/nature12912

Discovery and saturation analysis of cancer genes across 21 tumour types

Michael S. Lawrence[1], Petar Stojanov[1,2], Craig H. Mermel[1,3], James T. Robinson[1], Levi A. Garraway[1,2,4], Todd R. Golub[1,2,4,5], Matthew Meyerson[1,2,4], Stacey B. Gabriel[1], Eric S. Lander[1,4,6]* & Gad Getz[1,3,4]*

Although a few cancer genes are mutated in a high proportion of tumours of a given type (>20%), most are mutated at intermediate frequencies (2–20%). To explore the feasibility of creating a comprehensive catalogue of cancer genes, we analysed somatic point mutations in exome sequences from 4,742 human cancers and their matched normal-tissue samples across 21 cancer types. We found that large-scale genomic analysis can identify nearly all known cancer genes in these tumour types. Our analysis also identified 33 genes that were not previously known to be significantly mutated in cancer, including genes related to proliferation, apoptosis, genome stability, chromatin regulation, immune evasion, RNA processing and protein homeostasis. Down-sampling analysis indicates that larger sample sizes will reveal many more genes mutated at clinically important frequencies. We estimate that near-saturation may be achieved with 600–5,000 samples per tumour type, depending on background mutation frequency. The results may help to guide the next stage of cancer genomics.

Comprehensive knowledge of the genes underlying human cancers is a critical foundation for cancer diagnostics, therapeutics, clinical-trial design and selection of rational combination therapies. It is now possible to use genomic analysis to identify cancer genes in an unbiased fashion, based on the presence of somatic mutations at a rate significantly higher than the expected background level.

Systematic studies have revealed many new cancer genes, as well as new classes of cancer genes[1,2]. They have also made clear that, although some cancer genes are mutated at high frequencies, most cancer genes in most patients occur at intermediate frequencies (2–20%) or lower. Accordingly, a complete catalogue of mutations in this frequency class will be essential for recognizing dysregulated pathways and optimal targets for therapeutic intervention. However, recent work suggests major gaps in our knowledge of cancer genes of intermediate frequency. For example, a study of 183 lung adenocarcinomas[3] found that 15% of patients lacked even a single mutation affecting any of the 10 known hallmarks of cancer, and 38% had 3 or fewer such mutations.

In this paper, we analysed somatic point mutations (substitutions and small insertion and deletions) in nearly 5,000 human cancers and their matched normal-tissue samples ('tumour–normal pairs') across 21 tumour types. The questions that we examine here are: first, whether large-scale genomic analysis across tumour types can reliably identify all known cancer genes; second, whether it will reveal many new candidate cancer genes; and third, how far we are from having a complete catalogue of cancer genes (at least those of intermediate frequency). We used rigorous statistical methods to enumerate candidate cancer genes and then carefully inspected each gene to identify those with strong biological connections to cancer and mutational patterns consistent with the expected function.

The analysis reveals nearly all known cancer genes and revealed 33 novel candidates, including genes related to proliferation, apoptosis, genome stability, chromatin regulation, immune evasion, RNA processing and protein homeostasis. Importantly, the data show that the catalogue of cancer genes is still far from complete, with the number of candidate cancer genes still increasing sharply with sample size. These analyses enable us to estimate the sample sizes that will be needed to approach saturation.

Cancer-genome data

We collected and analysed data from 4,742 samples, consisting primarily of whole-exome sequence from tumour–normal pairs. The samples span 21 tumour types, which include 12 from The Cancer Genome Atlas (TCGA) and 14 from non-TCGA projects at the Broad Institute, with some overlapping tumour types (Table 1 and Supplementary Table 1). The number of samples per tumour type varied between 35 and 892.

Data were all analysed through the Broad's stringent filtering and annotation pipeline to obtain a uniform set of mutation calls (Methods). The data set consists of 3,078,483 somatic single nucleotide variations (SSNVs), 77,270 small insertions and deletions (SINDELs) and 29,837 somatic di-, tri- or oligonucleotide variations (DNVs, TNVs and ONVs, respectively), with an average of 672 per tumour–normal pair. The mutations include 540,831 missense, 207,144 synonymous, 46,264 nonsense, 33,637 splice-site, and 2,294,935 non-coding mutations (used to improve our background model). The analysis has sensitivity of >90% based on the sequencing depth and tumour purity and ploidy[4,5].

Mutation frequencies vary over more than five orders of magnitude (from 0.03 per megabase (Mb) to 7,000 per Mb) within and across tumour types, consistent with our recent study of mutational heterogeneity[6] of approximately 3,000 samples (of which 2,502 are included in this data set) (Supplementary Fig. 1). Mutation spectra also vary sharply within and across tumour types[6] (Supplementary Fig. 2).

Cancer-genome analysis

We analysed these data to identify candidate cancer genes, by which we mean genes harbouring somatic point mutations (that is, substitutions

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, Massachusetts 02215, USA. [3]Massachusetts General Hospital, Cancer Center and Department of Pathology, 55 Fruit Street, Boston, Massachusetts 02114, USA. [4]Harvard Medical School, 25 Shattuck Street, Boston, Massachusetts 02115, USA. [5]Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, Maryland 20815, USA. [6]Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.
*These authors contributed equally to this work.

Table 1 | List of the 21 tumour types analysed

Tumour type	No. of tumour–normal pairs	Median somatic mutation frequency (per Mb)	No. of significantly mutated genes	No. of additional significant genes found under RHT
Acute myeloid leukaemia	196	0.4	26	1
Bladder	99	7.1	24	10
Breast	892	1.2	32	5
Carcinoid	54	0.5	1	0
Chronic lymphocytic leukaemia	159	0.6	7	8
Colorectal	233	3.1	23	12
Diffuse large B-cell lymphoma	58	3.3	16	7
Endometrial	248	2.5	58	15
Oesophageal adenocarcinoma	141	4.0	8	7
Glioblastoma multiforme	291	2.2	22	4
Head and neck	384	3.9	25	9
Kidney clear cell	417	1.9	15	6
Lung adenocarcinoma	405	8.1	22	10
Lung squamous cell carcinoma	178	9.9	11	13
Medulloblastoma	92	0.3	2	1
Melanoma	118	12.9	19	9
Multiple myeloma	207	1.6	11	3
Neuroblastoma	81	0.5	1	0
Ovarian	316	1.7	5	5
Prostate	138	0.7	4	2
Rhabdoid tumour	35	0.1	1	0

The number of significantly mutated genes detected using the MutSig suite when analysing the full set of genes. RHT, restricted hypothesis testing on the set of cancer genes found in all the other tumour types. Supplementary Table 3 lists the cancer genes found in each tumour type and their frequencies (per cent of patients with mutations).

and small insertion or deletions) at a statistically significant rate or pattern in cancer. (Such genes will ultimately need to be verified by biological experiments to be considered validated cancer genes.) In this paper, we do not seek to implicate genes based on other criteria (such as amplification or deletion, translocations or epigenomic modification; however, see ref. 7 for an analysis of copy-number alterations across many tumour types).

In principle, candidate cancer genes can be discovered by sequencing enough tumour–normal pairs; based on the presence of an excess of somatic mutations compared to expectation. However, careful analysis is required to assess statistical significance. The mere presence of somatic mutations is insufficient to implicate a gene in cancer, inasmuch as 93% of genes carried mutations in at least five samples.

We showed recently[6] that heterogeneity of mutation rates and patterns in cancer can give rise to false positives and described methods to overcome this problem. We applied these methods to identify candidate cancer genes. We used the most recent version of the MutSig suite of tools (Supplementary Fig. 3a and Methods), which looks for three independent signals: high mutational burden relative to background expectation, accounting for heterogeneity; clustering of mutations within the gene[8]; and enrichment of mutations in evolutionarily conserved sites[8]. We combined the significance levels (*P* values) from each test to obtain a single significance level per gene (Methods).

We analysed each tumour type separately, as well as the entire cohort ('combined' set), using the same methodology to ensure that the results can be compared across types. We verified that each analysis accurately calculates the significance level of genes, based on the fact that the vast majority of genes fit the null hypothesis and lie on the diagonal in a Q–Q plot (Supplementary Fig. 3b). For each analysis, genes with false discovery rate (FDR) $q \leq 0.1$ were declared to be candidate cancer genes (Methods). Using an FDR of $q \leq 0.1$ ensures that the expected fraction of false positives in each analysis does not exceed 10%. This well-established statistical procedure results in an increase in statistical power to detect true positives, while controlling the proportion of false positives. We also analysed the merged set of gene × tumour-type pairs identified from the 22 individual analyses (here we include the combined set as one of the 'tumour types'), using methods discussed below.

Data and results are posted at http://www.tumorportal.org/. The site includes graphical displays of the mutations in each of the 18,388 genes studied; see examples in Fig. 1 and Supplementary Fig. 4. The site also includes tables of mutational data for each significant gene and Q–Q plots for each statistical test.

Candidate cancer genes across 21 tumour types

A total of 334 gene × tumour-type pairs were found by our analysis to be significantly mutated. These 334 pairs involve 224 distinct genes. The number of genes detected per tumour type varied considerably (range of 1–58), with 7 types having fewer than 10 genes and 2 (breast and endometrial) having more than 30 (Fig. 2, Supplementary Fig. 5 and Table 1). The specific genes differed substantially across tumour types, although some pairs of tumour types showed clear similarity, such as lung squamous cancer and head and neck squamous cancer (Methods and Supplementary Fig. 6).

Notably, only 22 genes were found to be significant in three or more tumour types. The well-established cancer genes *TP53, PIK3CA, PTEN, RB1, KRAS, NRAS, BRAF, CDKN2A, FBXW7, ARID1A* and *MLL2*, as well as *STAG2*, were significant in four or more tumour types. An additional 10 genes (*ATM, CASP8, CTCF, ERBB3, HLA-A, HRAS, IDH1, NF1, NFE2L2* and *PIK3R1*) were significant in three tumour types.

Although the power to detect cancer genes varied across tumour types (based on sample size and background mutation frequency), the marked differences across tumour types do not simply reflect differences in detection power. For example, tumour types with low mutation frequency or many samples often show fewer cancer genes despite having greater statistical power to detect them (Table 1). Moreover, many genes that are highly enriched in one (for example, *VHL, WT1*) or a few (for example, *HRAS, FBXW7*) tumour types fail to show detectable enrichment across the entire data set (Supplementary Table 2). Notably, most of the significant gene × tumour-type pairs involve only a small fraction of patients (with one half of the significant pairs involving ≤6.1% of patients, and one quarter involving ≤3.1%).

We then analysed the combined set, which produced 114 genes (Fig. 3 and Supplementary Table 2). Although 84 of these genes were already identified from analysis of individual tumour types, the remaining 30 achieved significance based only on the frequency of mutations across tumour types, underscoring the value of cross-tumour-type analysis. Conversely, 140 of the 224 genes found in analysis of individual tumour types did not reach significance when analysing the combined set (Fig. 3, bottom-right quadrant), consistent with the observation that many genes show strong enrichment in only one or a few tumour types.

By merging the 22 lists above, we obtained a Cancer5000 set containing 254 genes. Although the expected proportion of false positive genes in each list does not exceed 10%, the expected proportion in the merged list is actually higher (because true positives will tend to occur across several tumour types, whereas false positives will tend to be random

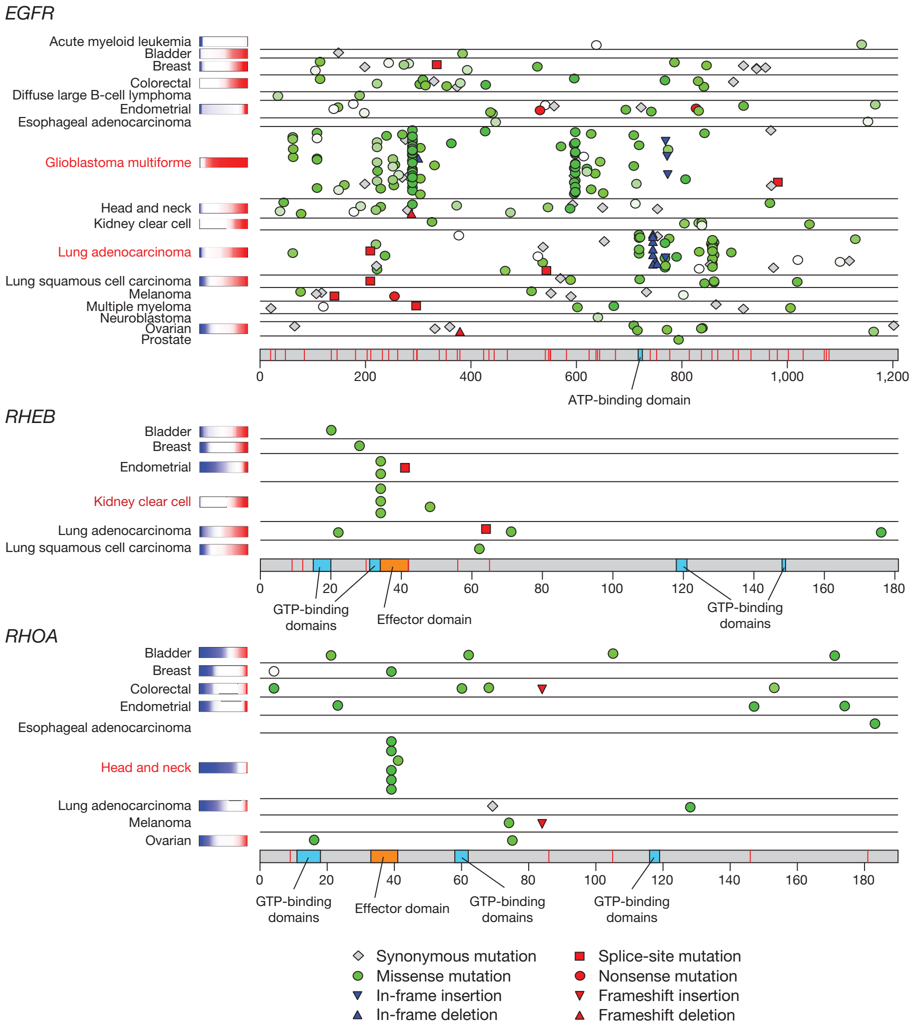


Figure 1 | Mutation patterns for one known and two novel cancer genes. *EGFR* shows distinctive tumour-type-specific concentrations of mutations in different regions of the gene. *RHEB*, which encodes a small GTPase in the Ras superfamily, shows a mutational hotspot in the effector domain. *RHOA*, another a member of the Ras superfamily, also shows a mutational hotspot in the effector domain. Coloured bars after tumour-type names are copy-ratio distributions for the gene, when available (red, amplified; blue, deleted). Missense mutations are represented by green circles of varying saturation indicating degree of evolutionary conservation of the mutated base pair, from highly conserved (dark green), to medium conservation (light green), to totally unconserved (white). Tumour types with names shown in red were significantly mutated in this gene, in dark red were nearly significantly mutated, or in black were not significantly mutated. Thin red strokes in the protein ideogram represent splice sites (see also Supplementary Fig. 4; similar diagrams for all genes are available at http://www.tumorportal.org).

singletons). A rigorous solution is to analyse the gene × tumour-type pairs as approximately 400,000 distinct hypotheses (approximately 18,400 genes × 22 types) and apply an FDR of $q \leq 0.1$. This analysis produces 403 significant pairs, which involve 219 distinct genes. We refer to this set as the Cancer5000-S (for 'stringent') genes. (All but six of the genes are contained in the Cancer5000 set.) Of the 403 significant pairs, 10% (approximately 40) at most are expected to be false positives. Assuming conservatively that the 40 pairs affect 40 distinct genes, we expect 179 of the 219 genes to be true cancer genes. Below, we discuss genes from both the Cancer5000 and Cancer5000-S sets.

Coverage of known cancer genes

We first asked whether all cancer genes that have been discovered and validated to date can be identified by hypothesis-free genomic analysis. As a reference set, we used the Cancer Gene Census (CGC), which is a manually curated catalogue of cancer genes. The current version (v65) contains 130 cancer genes driven by somatic point mutations (as well as additional genes mutated by other mechanisms), of which 82 are associated with 1 or more of the 21 tumour types studied here.

Of these 82 genes, 60 were identified in our Cancer5000 set. Of the remaining 22 genes, 8 fell just below significance in our data set, 6 appear in the CGC based on focused studies of the gene in very large samples (typically >1,000), and 8 genes harboured few mutations and seem to lack adequate evidence to justify association with any of the tumour types we studied. The first two categories would clearly be captured with larger sample sizes.

Analysis of novel candidate cancer genes

Of the 219 genes in the Cancer5000-S set, 81 are neither listed in the CGC as affected by point mutations in these tumour types (v65) nor discussed in papers published so far (Supplementary Table 4). (The list includes three genes that appear in tables in published papers based on mutations in a handful of samples, but were not noted or interpreted in the text.) Of the 41 additional genes in the Cancer5000 (but not Cancer5000-S) set, none are in the CGC but 3 are reported in recent publications (Supplementary Table 4).

We closely analysed these 81 'novel' genes to look for connections with cancer biology, together with a mutational pattern consistent with the biology. Where loss-of-function would be expected, we looked for an excess of disruptive changes, such as nonsense and frameshift mutations. In cases in which gain-of-function would be expected, we examined whether the overall collection of mutations included hotspots that resulted in recurrent changes at identical or nearby amino acids (often causing precisely the same change). Conversely, where we observed distinctive mutation patterns, we examined whether they were consistent with known biology.

As discussed above, the Cancer5000-S set is expected by design to contain approximately 40 false positives. Assuming conservatively that these false positives fall exclusively in the novel set, we expect approximately 41 of the 81 novel genes to be true positives.

In fact, we identified strong and consistent connections to cancer for at least 21 of the novel genes in the Cancer5000-S set. Among the 38 additional novel genes in the larger Cancer5000 set, we found 12 additional strong candidates. (References supporting the biological

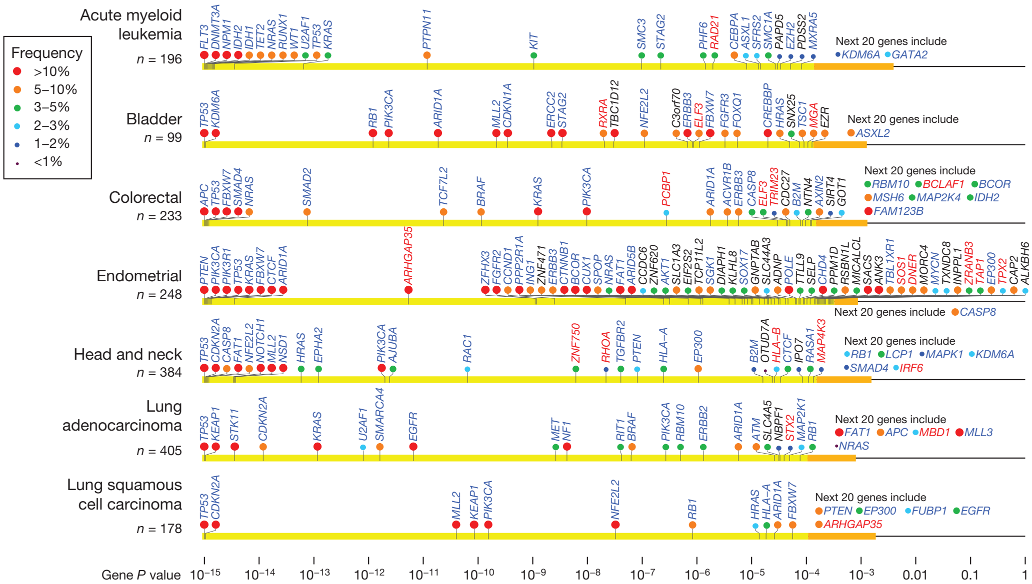


Figure 2 | Cancer genes in selected tumour types. Genes are arranged on the horizontal line according to *P* value (combined value for the three tests in MutSig). Yellow region contains genes that achieve FDR q ≤ 0.1. Orange interval contains *P* values for the next 20 genes. Gene-name colour indicates whether the gene is a known cancer gene (blue), a novel gene with clear connection to cancer (red, discussed in text), or an additional novel gene (black). Circle colour indicates the frequency (percentage of patients carrying non-silent somatic mutations) in that tumour type; see also Supplementary Fig. 5.

roles of the genes are provided in Supplementary Table 5.) We briefly describe below these 33 genes not previously reported as significantly mutated in cancer.

Four genes encode anti-proliferative proteins, in which loss-of-function mutations would be expected to contribute to oncogenesis. A notable example is *ARHGAP35* (previously called *GRLF1*), which encodes a Rho-GTPase-activating protein, for which only a single tumour type reaches statistical significance on its own, but which gives a strong signal ($q = 2 \times 10^{-12}$) in the combined set of 4,742 tumours (83 missense, 38 nonsense, 16 frameshift and 2 splice site). Notably, the gene resides in a small genomic region that is focally deleted in many tumours. Other examples are *MGA*, whose product competes with Myc for binding to Max and which resides in small focal deletions (containing ≤4 genes) in ovarian and various epithelial cancers; the interferon regulatory factor *IRF6*, which is known to have tumour suppressive roles in keratinocytes and is mutated in head and neck squamous cancer; and the delta/notch-like EGF-repeat gene *DNER*.

Six additional genes encode proteins that are clearly involved in cell proliferation: *RHEB*, *RHOA*, *SOS1*, *ELF3*, *SGK1* and *MYOCD*. Notably, *RHEB* and *RHOA* encode small GTPases, in which recurrent mutations affect the 9-amino-acid effector domain. For *RHEB*, five tumours (two endometrial and three kidney clear cell cancer) carry Tyr35Asn mutations, which alter the first amino acid of the effector domain. For *RHOA*, seven tumours (six head and neck, one breast) carry mutations affecting the effector domain: these include six Glu40Gln mutations and a single Tyr42Ile mutation, which alter the seventh and ninth amino acids of the effector domain, respectively. *SOS1* encodes a guanine nucleotide exchange factor that promotes activation of Ras proteins, in which gain-of-function mutations might contribute to oncogenesis. Consistent with this notion, *SOS1* carries Asn233Tyr mutations in six tumours (four endometrial and two lung adenocarcinoma) and Arg 552 alterations in three tumours (two endometrial and one AML). Notably, the same Arg 522 alterations in *SOS1* have been found to be germline mutations causing Noonan syndrome and to cause gain of function, resulting in Ras activation. *ELF3* encodes an ETS-domain transcription factor that functions in cell differentiation; it carries many truncating mutations in bladder and colon cancer. Myocardin (*MYOCD*), which encodes a transcriptional regulator involved in differentiation and cell migration, has a cluster of 9 mutations at amino acids 750–770 (7 in melanoma, 1 head and neck, 1 lung adenocarcinoma) with a hotspot of four at Ser 763. The retinoid X receptor alpha *RXRA*, which forms a heterodimer with retinoic acid receptors to regulate cell growth and survival, shows a clear hotspot of recurrent mutations at Ser 427 in bladder cancer and nearby mutations in lung, head and neck, and oesophageal cancers.

Five genes encode pro-apoptotic factors, in which loss-of-function mutations would be expected to promote oncogenesis. These genes encode alpha-kinase 2 (*ALPK2*); Bcl2-associated factor 1 (*BCLAF1*); a MAP kinase (*MAP4K3*) reported to post-transcriptionally regulate the apoptotic proteins PUMA (also known as BBC3), BAD and BIM (also known as BCL2L11); a zinc-finger protein (*ZNF750*, which harbours many early frameshift and nonsense mutations in head and neck cancer and is the only known gene residing in a small current focal deletion in head and neck and lung squamous cancers); and tumour necrosis factor (*TNF*, which harbours mutations in five diffuse large B-cell lymphomas that are tightly clustered in the region encoding the membrane and cytoplasmic domain, rather than the soluble TNF protein).

Six genes encode proteins related to genome stability. These include *CEP76* (encoding a centrosomal protein, whose depletion drives aberrant amplification of centrioles), which harbours early nonsense mutations in many tumour types and resides in a focal deletion peak in acute myeloid leukaemia; *RAD21* (encoding a protein crucial for chromosome segregation and double-strand break repair), which is mutated at significant rates in acute myeloid leukaemia and also harbours mutations in other tumour types; the p53-binding protein *TP53BP1* (encoding a check-point protein that binds to double-strand breaks), which does not reach significance in any single tumour type, but is significant in the combined data set owing to truncating mutations in many tumour types; *TPX2* (encoding a protein involved in mitotic spindle formation, whose depletion leads to aneuploidy); and *ZRANB3* (encoding a translocase that helps to rescue stalled replication forks). In addition, *STX2* encodes a protein required for cytokinesis, whose disruption may promote aneuploidy; *STX2* harbours recurrent mutations at Arg 107 in lung and endometrial tumours.

Five genes are associated with chromatin regulation. *SETDB1* encodes a H3K9 histone methyltransferase (*SETD2*, which encodes a H3K36 histone methyltransferase, has been shown previously to be mutated in cancer). *MBD1* encodes a protein that binds methylated-CpG and is required for SETDB1 activity; it contains five mutations in endometrial cancer in the amino-terminal methyl-binding domain. *EZH1* encodes a H3K27 histone methyltransferase; it does not reach significance in any individual tumour type, but is significant in the combined set owing to truncating mutations in multiple tumour types. *EZH1* shows a similar pattern of mutations as seen in the well-established cancer gene *EZH2*, with truncating mutations along the gene and a hotspot of

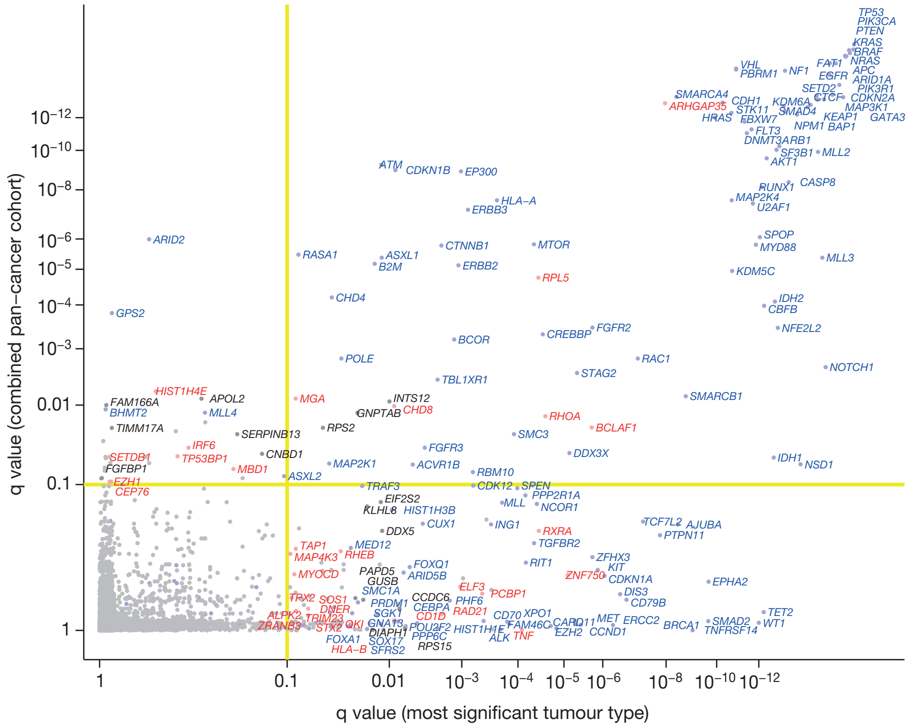


Figure 3 | Cancer genes identified from a data set of 4,742 tumours. Genes are plotted by the q value (FDR) in the most significant of the 21 tumour types (*x* axis) and the q value when the 4,742 tumours are analysed as a combined ('pan-cancer') cohort (*y* axis). Genes in the top-left quadrant reached significance only in the combined analysis. Genes in the bottom-right quadrant reached significance only in one or more single-type analyses. Genes in the top-right quadrant were significant in both the combined set and in individual tumour types. Colour of gene names is as in Fig. 2.

mutations within the SET domain. *CHD8* encodes a chromatin helicase DNA binding protein (like the known cancer gene *CHD4*) that suppresses the β-catenin–Wnt signalling pathway. The histone protein *HIST1H4E* is mutated in multiple tumour types; two other histone genes, *HIST1H1E* and *HIST1H3B*, have previously been reported as significantly mutated in CLL and DLBCL, respectively.

Three genes encode proteins whose loss is expected to help tumours evade immune attack; they all recurrently subject to truncating mutations across several tumour types. These include the major histocompatibility protein *HLA-B* (loss of the *HLA-A* gene has been implicated in lung cancer), *TAP1* (which processes intracellular peptides for presentation to the immune system) and *CD1D* (which presents lipid antigens to natural killer cells), the last of which shows a cluster of truncating mutations at the internalization domain that are likely to abolish antigen-presentation function.

Three genes are associated with RNA processing and metabolism. *PCBP1*, whose protein product blocks translation of certain mRNAs by binding to poly(C) regions of messenger RNAs, carries two mutations in each of two nearby leucines (Leu 100 and Leu 102) that mediate dimerization of the protein's K-homology domains. We speculate that disruption of *PCBP1* leads to increased translation of one or more pro-oncogenic mRNAs. *QKI* encodes an RNA-binding protein that regulates pre-mRNA splicing, including the known cancer gene *CDKN1B*; the gene harbours caboxy-terminal truncating mutations in several tumour types that are likely to remove the nuclear localization signal; and the gene resides in a recurrent deletion peak in glioblastoma and ovarian cancer. Finally, the ribosomal protein gene *RPL5* contains early truncating mutations in glioblastoma and other tumour types and resides in a focally deleted region in glioblastoma; heterozygous loss of certain ribosomal proteins has been reported to contribute to cancer.

One gene, *TRIM23*, is involved in protein homeostasis. It encodes an ubiquitin E3 ligase and harbours recurrent mutations at Asn 93 (four tumours) and Asp 289 (three tumours). Mutations in this gene may promote cancer by altering the substrate specificity of the E3 ligase in a manner that leads to accumulation of an oncogenic protein.

Beyond these 33 genes, the set of 81 novel genes is likely to contain additional true cancer genes. For example, we omitted genes with connections to cancer (such as *HSP90AB1*, *PPM1D* and *ITGB7*) in situations in which we could not easily reconcile the function in cancer with the observed pattern of mutations. In addition, we may have overlooked additional candidate cancer genes because we did not identify clear connections with cancer, owing to gaps in the literature or in our knowledge.

Saturation analysis

We next explored whether the discovery of candidate cancer genes is approaching saturation or whether many more genes are likely to be found. An effective test is to perform 'down-sampling'; that is, to study how the number of discoveries increases with sample size, by repeating the analysis on random subsets of samples of various smaller sizes.

For each tumour type (omitting those with five or fewer candidate cancer genes), the number of genes increases roughly linearly with sample size (examples in Fig. 4a; see also Supplementary Fig. 7), indicating that the inventory for each of the tumour types is far from complete. The number of genes also increases linearly with the number of tumour types studied (Fig. 4b), suggesting that it is valuable to increase both the sample size per tumour type and the number of tumour types.

We also studied how the total number of candidate cancer genes varies with sample size when applying the 'stringent' methodology used to create the Cancer5000-S set. Here too, the total number of genes increases steadily with sample size (Fig. 4c). Notably, the saturation analysis varies considerably with the mutation frequency (Fig. 4d). Genes mutated in >20% of tumours are approaching saturation; those mutated at frequencies of 10–20% are still rising rapidly, but at a decreasing rate; those at 5–10% are increasing linearly; and those at <5% are increasingly at an accelerating rate.

We next sought to infer the nature of the genes awaiting discovery in each tumour type. One possibility is that some of these genes are already contained in the Cancer5000 set (by virtue of their contribution to other tumour types) but have not yet reached statistical significance in the given tumour type due to insufficient sample size. To test this idea, we performed restricted hypothesis testing (RHT): for each tumour type T, we omitted that tumour type, determined the set of genes (G_T) that are significant based on the remaining tumour

types, and determined which genes in G_T reached significance in the omitted tumour type when correcting for multiple-hypothesis testing based on only the number of genes in G_T rather than all (approximately 18,400) genes in the genome.

The RHT analysis implicated many additional Cancer5000 genes in the individual tumour types (median 6 per tumour type, range 0–15). The number of significant gene $\times$ tumour-type pairs increased from 334 to 461 across the 21 tumour types. The RHT analysis indicates that, with somewhat larger sample size, these genes are likely to reach significance in an unrestricted test (Table 1 and Supplementary Table 3). For some tumour types, the number of implicated genes more than doubled: lung squamous cell carcinoma increased from 11 to 24; CLL from 7 to 15; and ovarian from 5 to 10. Notably, three genes now became significant in four tumour types each (*ARID2*, *ERBB2*, *ARHGAP35*) and seven genes in three types each (*CTNNB1*, *FGFR3*, *KRAS*, *PTEN*, *SMAD4*, *MLL3*). Although nine of these genes are well known cancer genes, one (*ARHGAP35*) is absent from the current CGC list. Notably, *ARHGAP35* appears in the Cancer5000 set because it is significantly mutated in endometrial cancer (although not discussed in the recent TCGA publication[9]), but our RHT analysis also finds it to be significant in lung adenocarcinoma, lung squamous cell carcinoma, kidney clear cell, and head and neck cancer. The genes found to be significant in additional tumour types in the RHT analysis are mutated at a median frequency of 3.4%.

However, the data also clearly show that many new candidate cancer genes remain to be discovered beyond those in the current Cancer5000 set. First, in addition to the Cancer5000 genes being shown by RHT to be significant in additional tumour types, the down-sampling analysis shows that the number of novel genes being identified is increasing sharply (using the stringent analysis used to create the Cancer5000-S set). Second, adding additional tumour types typically adds novel 'tumour-type-specific' genes, which are unique to (or at vastly higher frequency in) the tumour type.

Power analysis

As the cancer-gene catalogue remains far from complete, we explored what sample sizes are needed to approach saturation. The power to detect a gene as significantly mutated depends on the properties of the tumour type, namely the average background somatic mutation frequency along the genome for the tumour type ('noise'), and the target

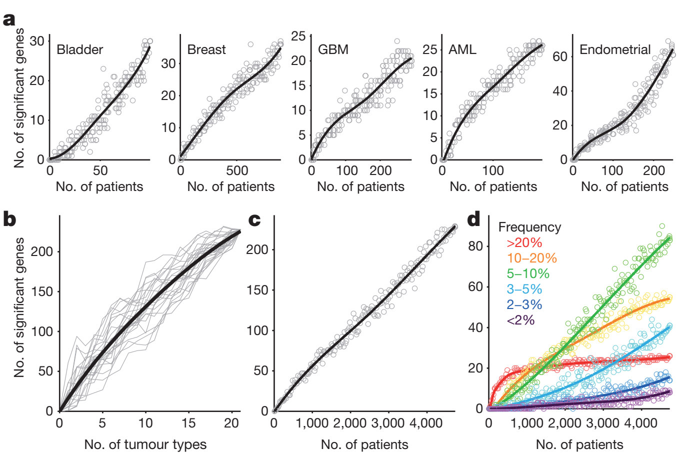


Figure 4 | Down-sampling analysis shows that gene discovery is continuing as samples and tumour types are added. **a**, Analysis within tumour types. Each point represents a random subset of patients. Line is a smoothed fit. **b**, Analysis by adding tumour types. Each grey line represents a random ordering of the 21 tumour types. **c**, Analysis by adding samples. Each point is a random subset of the 4,742 patients. **d**, Analysis in **c** broken down by mutation frequency. Genes mutated at frequencies $\geq$20% are nearing saturation, and intermediate frequencies show steep growth; see also Supplementary Figs 7 and 8.

frequency (across patients)—above the background rate—that one wishes to detect ('signal'). It also depends on the properties of the gene, namely its background mutation frequency relative to other genes (which depends on length and local mutation rate). We set a target of having 90% power to detect 90% of all genes. In addition, we allow for a false negative rate of 10% in detecting mutations, which increases the sample size by slightly more than 10%.

Figure 5 shows that the current collection lacks the desired power to detect genes mutated at 5% above the background rate for 17 of the 21 tumour types and even at 10% for 7 of the tumour types. These results are consistent with the down-sampling analysis showing that candidate cancer genes with frequency $\geq$20% are approaching saturation, whereas the number of candidate cancer genes at lower frequencies is continuing to grow rapidly with sample size.

Creating a reasonably comprehensive catalogue of candidate cancer genes mutated in $\geq$2% of patients will require between approximately 650 samples (for tumours with ~0.5 mutations per Mb, such as neuroblastoma) to approximately 5,300 samples (for melanoma, with 12.9 mutations per Mb).

Discussion

Precision medicine for cancer will ultimately require a comprehensive catalogue of cancer genes to enable physicians to select the best combination therapy for each patient based on the cellular pathways disrupted in their tumour and the specific nature of the disruptions. Such a catalogue will also guide therapeutic development by identifying druggable targets. In addition, the catalogue and its underlying data will facilitate the interpretation of cell lines, animal models and clinical observations and will reveal patterns of co-occurrence, mutual exclusivity and lineage restriction, which may provide mechanistic insights with profound therapeutic implications.

Although a handful of cancer genes are mutated at high frequency, most cancer genes mutated in most patients occur at intermediate frequencies (2–20%). To provide therapeutic options for most patients, it will therefore be critical to identify and understand the pathway-level implications of all genes mutated at intermediate frequencies (2–20%).

With growing data sets across many tumour types, pan-cancer analyses are becoming of great interest[10,11]. In this paper, we studied somatic point mutations in a collection of nearly 5,000 tumour-normal pairs across 21 cancer types. We identified a Cancer5000 set containing 254 genes, based on merging results from each tumour type and the combined set, and a stringent Cancer5000-S set containing 219 genes, accounting for multiple-hypothesis testing across the types. Nearly all previously known cancer genes in these tumour types are contained within these sets or just below statistical significance.

After eliminating genes reported in the CGC or recent papers and accounting for the expected number of false positives, the stringent Cancer5000-S set is expected to contain approximately 41 novel candidate cancer genes, with additional candidate cancer genes expected in the larger Cancer5000 set. After close inspection, we found 33 genes (21 in the stringent set and 12 more in the larger set) with strong functional connections to cancer and mutation patterns consistent with the presumed function. These genes fall within known 'hallmarks' of cancer[3,12], including cell proliferation, apoptosis, genome stability, chromatin regulation, immune evasion, RNA processing and protein homeostasis. Follow-up studies will be required to confirm and understand the functional impact of the mutations in these genes.

Beyond identifying new candidate cancer genes, our study demonstrates that we are far from having a complete catalogue of cancer genes, with many genes at clinically important frequencies within individual tumour types and across cancer as a whole still awaiting identification. The number of such genes is still increasing steeply with the number of samples and the number of tumour types studied. Importantly, these new candidate cancer genes are not rare. Substantial ongoing increases are seen in each of the 10–20%, 5–10% and 2–5% ranges (Fig. 4d).

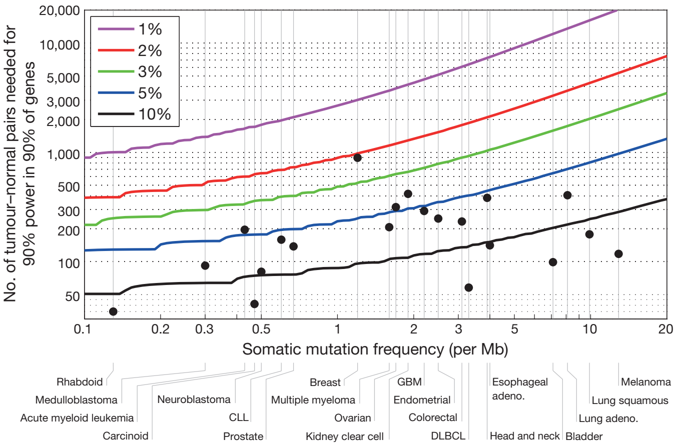


Figure 5 | Number of samples needed to detect significantly mutated genes, as a function of a tumour type's median background mutation frequency and a cancer gene's mutation rate above background. The number of samples needed to achieve 90% power for 90% of genes (y axis). Grey vertical lines indicate tumour type median background mutation frequencies (x axis). Black dots indicate sample sizes in the current study. For most tumour types, the current sample size is inadequate to reliably detect genes mutated at 5% or less above background; see also Supplementary Fig. 9. Adeno., adenocarcinoma.

Notably, of the 33 novel genes above, 5 are mutated at frequencies greater than 10% and fifteen at frequencies greater than 5%.

Creating a comprehensive catalogue of genes in which somatic point mutations propel cancer at both high (>20%) and intermediate (2–20%) frequency will require analysing an average of approximately 2,000 tumours for each of at least 50 tumour types, corresponding to approximately 100,000 tumours. (Currently defined tumour types may be divided, based on genomic information, into distinct subtypes, each of which should be analysed on its own. The ultimate number of tumour types will thus be defined iteratively by molecular analysis.)

Analysis should include both point mutations (as studied here), as well as other types of functional variation[7]. Genomic studies of such large numbers of samples is no longer prohibitive, in light of the one-million-fold decrease in the cost of DNA sequencing over the past decade. Given the devastating toll of cancer, with nearly 8 million deaths annually worldwide[13], completing the genomic analysis of this disease should be a biomedical imperative.

METHODS SUMMARY

For TCGA tumour types, mutation data were downloaded from the Synapse website. For non-TCGA tumour types, sequencing data was downloaded from dbGaP and processed through Firehose, the Broad Institute's analysis pipeline. Liftover was used to convert hg18 data. Each mutation in the combined MAF file was filtered against a panel of normal samples. Three significance metrics were calculated for each gene, using the previously described methods MutSigCV, MutSigCL, and MutSigFN. These measure the significance, respectively, of mutation burden, clustering, and functional impact. The three MutSig tests were combined into a single final P value for each gene, and q values were calculated using the method of Benjamini and Hochberg, and genes with $q \leq 0.1$ were declared to be candidate cancer genes. Down-sampling was performed within each tumour type and for the combined data set. MutSig analysis was repeated for a set of many smaller random subsets of patients. Genes were stratified by their maximal frequency across tumour types (Fig. 4d). Power analysis was performed using a binomial power model. We first calculated the probability, p_0, that a patient will have at least one non-silent mutation in a particular gene from the background model. We then calculated the signal we want to detect, $p_1 = p_0 + r(1-m)$, where r is the frequency of non-silent mutations in the population (above background) that a gene is mutated and m is the mis-detection rate of the mutation (we took $m = 0.1$). The power was then calculated using a binomial model, with $p = p_0$ representing the null hypothesis, and $p = p_1$ representing the alternative hypothesis (Methods). To obtain Fig. 5 and Supplementary Fig. 9 we found the number of tumour–normal pairs that yielded 90% power for 90% of genes as a function of background mutation frequency and r.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 September; accepted 27 November 2013.
Published online 5 January 2014.

1. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153,** 17–37 (2013).
2. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).
3. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150,** 1107–1120 (2012).
4. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnol.* **30,** 413–421 (2012).
5. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnol.* **31,** 213–219 (2013).
6. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
7. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature Genet.* **45,** 1134–1140 (2013).
8. Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl Acad. Sci. USA* **109,** 3879–3884 (2012).
9. Cancer Genome Atlas Research. Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
10. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).
11. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3,** 2650 (2013).
12. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
13. Ferlay, J. *et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer* **127,** 2893–2917 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was conducted as part of TCGA, a project of the National Cancer Institute and the National Human Genome Research Institute. We are grateful to T. I. Zack, S. E. Schumacher, and R. Beroukhim for sharing their copy-number analyses before publication.

Author Contributions G.G., E.S.L., T.R.G., M.M., L.A.G. and S.B.G. conceived the project and provided leadership. M.S.L., G.G., E.S.L., P.S. and C.H.M. analysed the data and contributed to scientific discussions. M.S.L., E.S.L. and G.G. wrote the paper. J.T.R., M.S.L., E.S.L. and G.G. created the website for visualizing this data set.

Author Information The data analysed in this manuscript have been deposited in Synapse (http://www.synapse.org), accession number syn1729383, and in dbGaP (http://www.ncbi.nlm.nih.gov/gap), accession numbers phs000330.v1.p1, phs000348.v1.p1, phs000369.v1.p1, phs000370.v1.p1, phs000374.v1.p1, phs000435.v2.p1, phs000447.v1.p1, phs000450.v1.p1, phs000452.v1.p1, phs000467.v6.p1, phs000488.v1.p1, phs000504.v1.p1, phs000508.v1.p1, phs000579.v1.p1, phs000598.v1.p1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.S.L. (lander@broadinstitute.org) and G.G. (gadgetz@broadinstitute.org).

METHODS

Mutation data and preprocessing. Mutation data were obtained as follows. For TCGA tumour types, mutation data were downloaded from the Synapse website (http://www.synapse.org), accession no. syn1729383. For non-TCGA tumour types, sequencing data was downloaded from dbGaP and processed through Firehose, the Broad Institute's analysis platform (http://www.broadinstitute.org/cancer/cga/Firehose). For tumour types that were originally aligned to build hg18, LiftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver) was used to convert the coordinates of each mutation to build hg19. All mutation data were then combined into a single MAF file. Duplicate patients and duplicate mutations were removed. To standardize the definition of a 'splice-site' mutation, any mutation affecting the two bases before or after a splice junction, was labelled as a splice-site mutation. Filtering was performed as follows. To remove common sequencing artefacts or residual germline variation, each mutation in the combined MAF file was subjected to a 'Panel of Normals' filtering process using a panel of over 4000 BAM files from normal samples. For each mutation, the position of the mutation was examined in each normal BAM file. Mutations observed in the panel of normals were removed from the MAF. The final MAF is available at http://www.tumorportal.org/.

MutSig significance calculations. Three significance metrics were calculated for each gene, using the previously described methods MutSigCV, MutSigCL, and MutSigFN. These measure the significance of mutation burden, clustering, and functional impact, respectively (Supplementary Fig. 3). MutSigCV was described previously[6]. MutSigCV determines the P value for observing the given quantity of non-silent mutations in the gene, given the background model determined by silent (and noncoding) mutations in the same gene and the neighbouring genes of covariate space that form its 'bagel'. MutSigCL and MutSigFN were used previously[8] but were not given names in that work. Here we name the methods to reflect the type of evidence of positive selection that they are designed to detect. MutSigCL and MutSigFN measure the significance of the positional clustering of the mutations observed, as well as the significance of the tendency for mutations to occur at positions that are highly evolutionarily conserved (using conservation as a proxy for probably functional impact). MutSigCL and MutSigFN are permutation-based methods and their P values are calculated as follows: The observed nonsilent coding mutations in the gene are permuted T times (to simulate the null hypothesis, $T = 10^8$ for the most significant genes), randomly reassigning their positions, but preserving their mutational 'category', as determined by local sequence context. We used the following context categories: transitions at CpG dinucleotides, transitions at other C–G base pairs, transversions at C–G base pairs, mutations at A–T base pairs, and indels. Indels are unconstrained in terms of where they can move to in the permutations. For each of the random permutations, two scores are calculated: S_{CL} and S_{FN}, measuring the amount of clustering and function impact (measured by conservation) respectively. S_{CL} is defined to be the fraction of mutations occurring in hotspots. A hotspot is defined as a 3-base-pair region of the gene containing many mutations: at least 2, and at least 2% of the total mutations. S_{FN} is defined to be the mean of the base-pair-level conservation values for the position of each non-silent mutation, as obtained from an alignment of 45 vertebrate genomes to the human genome, the UCSC 'phyloP46way' track, which can be downloaded from (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/README.txt). To determine a P_{CL}, the P value for the observed degree of positional clustering, the observed value of S_{CL} (computed for the mutations actually observed), was compared to the distribution of S_{CL} obtained from the random permutations, and the P value was defined to be the fraction of random permutations in which S_{CL} was at least as large as the observed S_{CL}. The P value for the conservation of the mutated positions, P_{FN}, was computed analogously. Finally, we noted that the gene *AJUBA* was referred to in some analyses by the alternative name *JUB*; after reconciling this naming difference, the gene was significant and added to the list of significant genes.

Combining MutSig statistics. The three MutSig tests described above (MutSigCV, MutSigCL and MutSigFN) were combined into a single final P value as follows. First, a joint P value (CL + FN) for the observed clustering and conservation was calculated from the joint probability distribution of the random permutations. Next, this was combined with the MutSigCV P value using two methods: the Fisher method of combining P values from independent tests (http://en.wikipedia.org/wiki/Fisher's_method); the truncated product method (TPM) for combining P values, which rewards highly significant P values in any one of the tests The combined P values for both methods were extremely similar. We examined the performance of each of the three metrics separately and each pairwise combination of two metrics. The results of these analyses are presented in Supplementary Table 1 (last tab) and summarized in Supplementary Table 5.

Multiple hypothesis corrections. In the analysis of each tumour type, a total of 18,388 genes were analysed. To correct for these multiple hypotheses, the final MutSig P values were converted to FDR (q values) using the method of Benjamini and Hochberg, and genes with $q \leq 0.1$ were declared to be significantly mutated. This was also done for the analysis of the combined cohort. Genes with $q \leq 0.1$ in any tumour-type analysis or in the combined-cohort analysis were declared to be a member of the Cancer5000 list of candidate cancer genes.

To correct for the 22 analyses thus combined (corresponding to 22 chances for each gene to become significant), a further level of multiple hypothesis correction was applied. A list was made of the 18,388 genes $\times$ 22 analyses = 404,536 hypotheses. The Benjamini–Hochberg method was applied to this full set, yielding new FDR q values. Any gene involved in these gene $\times$ tumour-type pairs was declared to be a member of the stringently corrected Cancer5000-S list of genes.

Down-sampling analyses. To analyse the dependence of the number of significantly mutated cancer-associated genes upon the size of the data set being analysed, down-sampling was performed. Three different down-sampling analyses are described: first, down-sampling within each tumour type; second, down-sampling of the number of different tumour types; and third, down-sampling of the full Cancer5000-S procedure.

Down-sampling within each tumour type (Fig. 4a and Supplementary Fig. 7): for each tumour type, the MutSig analysis was repeated for a set of many smaller subsets of patients from that tumour type. The sizes of the subsets were chosen to sample regularly the interval from zero patients to the final total number of patients that were in the full analysis. For each of the random subsets thus defined, we repeated the full MutSig calculation (MutSigCV + MutSigCL + MutSigFN) and combined the results of the three tests as described above. This enabled us to determine which genes remained significant when analysing each smaller subset. We counted how many of the genes remained significant at each smaller set size, and plotted this number as a smoothed function of set size. This allowed us to demonstrate that the number of significantly mutated genes detected is continuing to rise steeply in each tumour type. We also repeated this same analysis for the full combined data set (4,742 patients), with similar results.

Down-sampling of the number of different tumour types (Fig. 4b). To examine the effect of adding whole tumour types, we performed the following analysis. We constructed 25 random orderings of the 21 tumour types, and for each ordering we constructed 20 subsets by sequentially adding whole tumour types according to that ordering. Then we repeated the whole MutSig analysis for each of these subsets. This produced a set of curves showing how the number of significantly mutated genes increased as a function of the number of tumour types included in the analysis. The curve depended on the exact ordering of the tumour types as they were added, but all curves showed a steady increase in the number of genes, even at the highest numbers of tumour types. This demonstrated the importance of continuing to sample additional tumour types. We also repeated the analysis with subtraction of the expected number of false positives (Supplementary Fig. 8a); the results were qualitatively unchanged.

Down-sampling of the full Cancer5000-S procedure (Fig. 4c): we repeated our procedure of constructing the Cancer5000-S list by applying the stringent procedure of correction for the approximately 400,000 hypothesis (18,388 genes $\times$ 22 analyses), and computed how many genes remained significant at each smaller set size. We plotted the number of significantly mutated genes detected as a function of set size. This produced a curve similar to down-sampling within each tumour type, with the number of significant genes continuing to rise steeply even at the largest set sizes. We also repeated the analysis with subtraction of the expected number of false positives (Supplementary Fig. 8b); the results were qualitatively unchanged. Furthermore, we stratified the genes according to their frequency (calculated as the maximal frequency across tumour types), and plotted separate curves for each of the following frequency categories: 20% and above, 10–20%, 5–10%, 3–5%, 2–3%, and below 2%. This clearly demonstrated that the 20% and above genes have largely been discovered. In contrast, genes at lower frequencies are continuing to be discovered (Fig. 4d). Note that rerunning the analysis produces slightly different results in every run since the calculation of P values has a stochastic component. The genes at the edge of significance (that is, ones with q value close to 0.1) may be declared as significant or insignificant with respect to the cut-off of $q = 0.1$ in different analyses. This slight fluctuation is standard for permutation-based methods.

Power calculations. Power analysis was performed using a binomial power model. We first calculated the probability, p_0, that a patient will have at least one non-silent mutation in a particular gene from the background model. The calculation is based on the length of the gene, L (in coding bases), the background mutation frequency, μ (in mutations per base), the gene-specific mutation rate factor, f_g, (calculated by MutSigCV), the 3:1 typical ratio of non-silent to silent mutations; $p_0 = 1 - (1 - \mu f_g)^{(3L/4)}$. We used $L = 1{,}500$, and $f_g = 3.9$ (representing the 90th percentile of $f_g L_g/1{,}500$ across the approximately 18,000 genes and $f_g = 1$ for the 50th percentile gene). We then calculated the signal we want to detect, $p_1 = p_0 + r(1 - m)$, where r is the frequency of non-silent mutations in the population (above background) that a gene is mutated and m is the mis-detection rate of the mutation (we took $m = 0.1$). The power was then calculated by: first,

using a binomial of N trials (that is, N patients) and $p = p_0$, finding the minimal number of patients with mutations that reach a genome-wide significance level ($P \leq 5\times10^{-6}$); and second, calculating the power as the probability of observing at least this many patients with mutations when using a binomial with $p = p_1$. To obtain Fig. 5 and Supplementary Fig. 9 we found the values of N that yielded 90% power as a function of μ and r.

ARTICLE

doi:10.1038/nature12893

Immunological and virological mechanisms of vaccine-mediated protection against SIV and HIV

Mario Roederer[1], Brandon F. Keele[2], Stephen D. Schmidt[1], Rosemarie D. Mason[1], Hugh C. Welles[1,3], Will Fischer[4], Celia Labranche[5], Kathryn E. Foulds[1], Mark K. Louder[1], Zhi-Yong Yang[1]†, John-Paul M. Todd[1], Adam P. Buzby[6], Linh V. Mach[6], Ling Shen[6], Kelly E. Seaton[7], Brandy M. Ward[5], Robert T. Bailer[1], Raphael Gottardo[8], Wenjuan Gu[9], Guido Ferrari[5], S. Munir Alam[7], Thomas N. Denny[7], David C. Montefiori[5], Georgia D. Tomaras[7], Bette T. Korber[4], Martha C. Nason[9], Robert A. Seder[1], Richard A. Koup[1], Norman L. Letvin[6]‡, Srinivas S. Rao[1], Gary J. Nabel[1]† & John R. Mascola[1]

A major challenge for the development of a highly effective AIDS vaccine is the identification of mechanisms of protective immunity. To address this question, we used a nonhuman primate challenge model with simian immunodeficiency virus (SIV). We show that antibodies to the SIV envelope are necessary and sufficient to prevent infection. Moreover, sequencing of viruses from breakthrough infections revealed selective pressure against neutralization-sensitive viruses; we identified a two-amino-acid signature that alters antigenicity and confers neutralization resistance. A similar signature confers resistance of human immunodeficiency virus (HIV)-1 to neutralization by monoclonal antibodies against variable regions 1 and 2 (V1V2), suggesting that SIV and HIV share a fundamental mechanism of immune escape from vaccine-elicited or naturally elicited antibodies. These analyses provide insight into the limited efficacy seen in HIV vaccine trials.

Among the five human efficacy trials of HIV-1 vaccines, only one has shown some success in preventing HIV infection. In the RV144 trial, a combination viral vector and protein immunization achieved a modest 31% efficacy in a cohort of low-risk adults in Thailand[1]. In-depth immunological correlates analysis suggested that specific antibody responses to the HIV-1 envelope variable regions 1 and 2 (V1V2) region correlated with protection, whereas an immunoglobin A (IgA) response showed a negative association[2,3]. Virus sequencing of the breakthrough infections in RV144 suggested a possible vaccine-mediated selection pressure against certain virus variants[4]; the mechanism of immune pressure remains elusive, but may include elicitation of antibodies targeting V1V2 of those variants[5]. In contrast, the recent HVTN 505 trial, using a DNA-prime, recombinant adenovirus type 5 (rAd5) boost, was halted for futility with no vaccine efficacy[6].

Infection of nonhuman primates with SIV represents the best available animal model for testing vaccine concepts for protecting against HIV infection, and mucosal challenge with SIV can be used to model human mucosal HIV exposure[7]. Several SIV challenge studies have shown partial protection from acquisition; in some cases, there has been an association to elicited antibodies, but a strong immunological mechanism or correlate has not been identified[8–13]. Here, we used a repetitive intrarectal challenge using a SIV_{smE660} challenge virus that was unmatched to the vaccines[14]. The E660 virus swarm is heterogeneous, comprising groups or clusters of viruses ranging from neutralization sensitive to resistant[15]. We reasoned that, in the absence of complete protection, the naturally occurring diversity of neutralization profiles would provide the most informative correlates analysis.

Our goals were to define cellular and humoral immune correlates of immunity, and to understand the mechanism leading to protection against SIV infection. Our immunogens included 'T-cell mosaics' designed to optimize coverage of epitope diversity for cellular responses[16,17]. We designed a four-arm study to define mechanisms of vaccine protection: (1) mosaic Gag; (2) mosaic heterologous envelope (Env); (3) heterologous Env based on a natural SIV_{mac239} sequence; and (4) control vaccine. Our primary questions were whether Env immunization is sufficient and/or necessary to provide protection against acquisition, whether Gag (alone) immunization provide any protection against acquisition, and finally whether the use of 'T-cell mosaic' Envs provide additional benefit over a natural Env sequence.

The number of acquisition end points in this study was similar to a large human efficacy study. We demonstrated that an Env-elicited immune response is necessary and sufficient to provide protection from acquisition. Importantly, by integrating immunological and virological analyses, we elucidated antibody-mediated mechanisms of protection and discovered a fundamental mechanism of virus escape from antibody-mediated control, shared by SIV and HIV, that has broad implications for understanding vaccine-mediated protection and potentially for vaccine design.

Vaccine immunogenicity

80 Indian origin rhesus macaques were enrolled in a DNA-prime, rAd5 boost immunization study. Animals were randomized into four groups of 20 based on *TRIM5α* (also known as *TRIM5*) alleles, gender, age and weight. All animals received three shots of DNA at 4-week intervals, followed by rAd5 at week 30[14]. The control group received vectors that contained no inserts; the second group ('mosaic Gag') received two SIV Gag mosaic immunogens[17]; the third group ('mosaic

[1]Vaccine Research Center, NIAID, NIH, Bethesda, Maryland 20892, USA. [2]SAIC-Frederick, Frederick National Laboratory, NIH, Frederick, Maryland 21702, USA. [3]George Washington University, Washington DC 20052, USA. [4]Los Alamos National Laboratories, Los Alamos, New Mexico 87545, USA. [5]Department of Surgery, Duke University, Durham, North Carolina 27710, USA. [6]Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02115, USA. [7]Human Vaccine Institute, Duke University, Durham, North Carolina 27710, USA. [8]Fred Hutchison Cancer Research Center, Seattle, Washington 98109, USA. [9]Biostatistics Research Branch, NIAID, NIH, Bethesda, Maryland 20892, USA. †Present address: Sanofi-Pasteur, Cambridge, Massachusetts 02139, USA. ‡Deceased.

Env') received two SIV Env mosaic immunogens (78% and 87% sequence identity to SIV_{smE543}, a clone similar to E660[16]); and the fourth group ('mac239 Env') received an immunogen encoding SIV_{mac239} Env (83% sequence identity to E543). Envelope sequences are shown in Supplementary Table 1, and sequence distances in Supplementary Table 2.

Vaccination elicited the expected cellular (Extended Data Fig. 1) and humoral (Extended Data Fig. 2) responses. Notably, compared to mac239 Env immunization, mosaic Env induced modestly lower and qualitatively different humoral responses (Extended Data Fig. 2). Mapping of the antibody response to unglycosylated linear peptides (Extended Data Fig. 2c) revealed that mac239 Env elicited a broader response than mosaic Env. Overall, immunization elicited mild neutralization and antibody-dependent cellular cytotoxicity activity against a limited set of viral strains (Extended Data Fig. 2d–g).

SIV challenge outcome

To test vaccine efficacy against infectious challenge, we exposed animals weekly to intrarectal administration of E660 at a dose that infects ~30% of control animals per exposure[14]. Each animal was challenged up to 12 times or until it had detectable plasma viraemia. Immunization with mac239 Env provided significant protection against acquisition, whereas mosaic Env immunization did not achieve significance (Fig. 1a). There was no difference in acquisition between Gag-immunized animals and control animals. For protection against acquisition, vaccine efficacy (V_E: the reduction in the rate of infection at each challenge)[18,19] was 69% for mac239 Env (Fig. 1d).

All infected animals that received active immunization showed 0.7 to 1.1 $\log_{10}$ decrease in peak viral load (VL) on average (Fig. 1b, c and Extended Data Fig. 3b). The best control of acute VL occurred in the mosaic Env arm, whereas the mosaic Gag arm showed the best long-term control (Fig. 1d). We confirmed previous findings that animals with certain alleles of *TRIM5α* showed better innate control of infection and pathogenesis[14] (Extended Data Fig. 3d, e). Due to the stratification by *TRIM5α* alleles in our study, including this genotype as a covariate in analyses does not affect our conclusions. All three vaccine arms showed protection against loss of CD4 cells (Extended Data Fig. 3c). Thus, the mosaic Env constructs elicited effective T-cell responses that protected against pathogenic effects of infection, despite their inability to block acquisition.

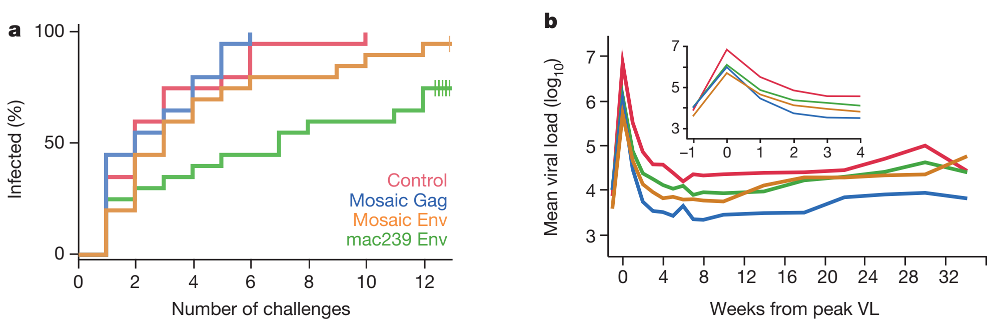
Transmitted founder analysis

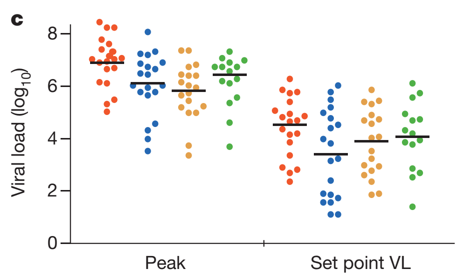
Because E660 is a viral swarm with 1.8% sequence diversity, the number of transmitted founder (T/F) viruses can be determined by single-genome amplification (SGA). For every infected animal, sequencing was done on plasma from the earliest time point with detectable plasma VL, 1 week after infection: thus, the inferred sequences represent the original infecting viruses[7]. Both Env arms showed a significant decrease in T/F variants (Fig. 2a). From these data, an efficacy can be calculated by the reduction of T/F variants per challenge; theoretically, this value estimates V_E for a very low (clinically relevant) infectious dose. Immunization with mac239 Env reduced the number of T/F variants by 81%, and mosaic Env reduced T/F by 51% (Fig. 2b).

Phylogenetic analysis using all complete Env sequences did not reveal an obvious clustering of T/F variants by vaccine arm. However, a strong 'sieving' effect was discerned by examining individual amino acid variants. Over the Env coding sequence, the 133 T/F sequences showed variation at 63 sites (Supplementary Table 3); 20 positions in the cytoplasmic domain or with rare variation (<5 of 133 T/F) were excluded from further analysis. Among the remaining sites, we found significant differences in variant representation in the Env vaccinated arms compared to the control and Gag arms (Fig. 2c and Extended Data Fig. 4). The strongest effect was seen at positions 23, 45 and 47. The consensus T/F sequence at these positions (VTR) was found in a majority of T/F viruses in the control and Gag arms. In contrast, variant sequences (IAK) were significantly overrepresented in the Env-immunized arms. Thus, immunization with Env sequences induced an immune response selecting against virions with the VTR signature.

Mechanism of virus selection in vaccinees

To define the mechanism of vaccine-mediated selection against viral variants, we measured the neutralization profile of all 40 Env-immunized animals against pseudo-typed viruses. CP3C-P-A8 ('CP3C' for brevity), a clone from the E660 swarm, is a neutralization-sensitive virus and has the amino acids VTR at positions 23, 45 and 47. CR54-PK-2A5 ('CR54'),





d

	Infection rate	Vaccine efficacy (*P*)	Viral load ($\log_{10}$) Peak (*P*)	Set point (*P*)
Control	34%		6.8	4.4
Mosaic Gag	38%	–12% (0.62)	6.1 (0.017)	3.4 (0.04)
Mosaic Env	22%	35% (0.12)	5.7 (0.001)	3.8 (0.16)
mac239 Env	11%	69% (0.0002)	6.1 (0.035)	4.0 (0.40)

Figure 1 | Protection against SIV challenge. **a**, The fraction infected animals in each arm following each of 12 challenges is shown. Five animals in the mac239 Env arm and one animal in the mosaic Env arm remained uninfected after 12 challenges. **b**, For each arm, the geometric mean plasma viral load (RNA copies per ml) for infected animals is shown. Each animal is synchronized to its peak VL. Inset, expanded scale for the acute phase. **c**, The peak and set point plasma viral load distributions for all infected animals. **d**, The infection rate is the fraction of infections out of the total number of exposures; vaccine efficacy was calculated as described in the methods.

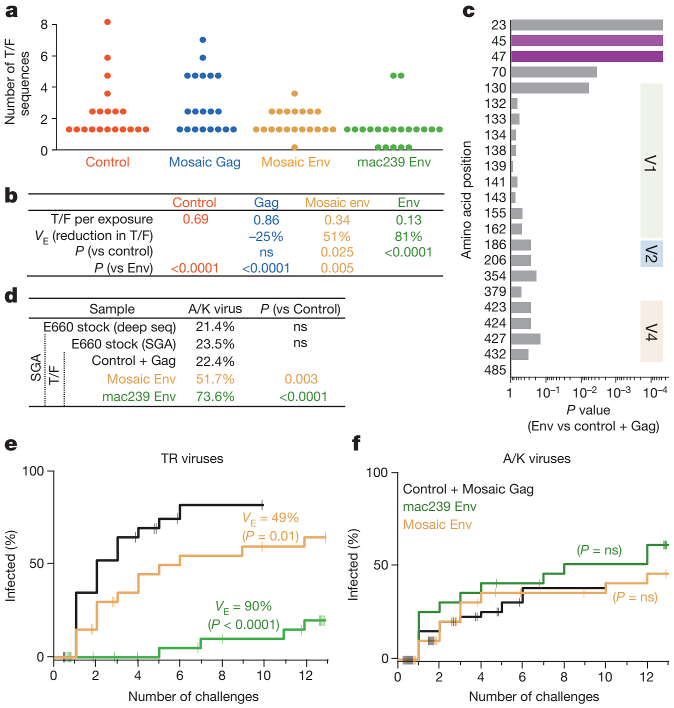


Figure 2 | Analysis of transmitted/founder (T/F) viruses. a, The distribution of unique T/F viruses in the first virus-positive plasma sample is shown for all 80 animals. **b**, The average number of T/F viruses per exposure event was calculated. Here, vaccine efficacy (V_E) is computed as the reduction in the number of T/F viruses (ns (not significant), $P > 0.05$). **c**, For each position in Env, the *P* value is shown for a permutation test comparing the fraction of viruses with the consensus amino acid in the Env T/F vs the control and Gag T/F. *P* values at positions 23, 45 and 47 remain significant after correction for multiple comparisons. **d–f**, Based on the sequence at positions 45 and 47, T/F viruses were divided into 'TR' (45T+47R) and 'A/K' (45A or 47K) viruses. **d**, Proportion of A/K viruses in the E660 challenge stock was measured by deep sequencing or by SGA, and among T/F in the immunization arms by SGA. A Fisher's exact test was performed to determine the significance of the difference in A/K viruses compared to the Control+Gag arms (ns, $P > 0.05$). **e**, **f**, Cumulative infection probabilities by TR or A/K viruses was done using a non-parametric estimate for competing risks[23]; the V_E and *P* values are computed using likelihoods from a modified Hudgens and Gilbert leaky vaccine model[18]. Tick marks indicate censoring of animals solely infected by the other virus type (challenges 1–12), or remaining uninfected after 12 challenges.

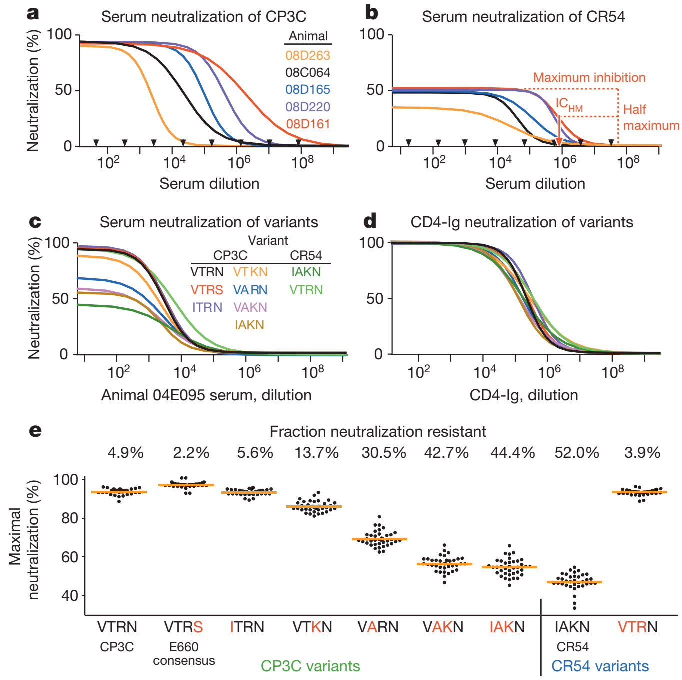


Figure 3 | Sequences accounting for neutralization resistance. a, **b**, Neutralization curves of CP3C, a sensitive clone from E660 (**a**), and of CR54, a resistant clone from E660 (**b**), using dilutions of sera from five Env-immunized animals (selected to show the range of potency). Black arrows indicate which dilutions were tested in duplicate; curves represent nonlinear least squares regressions of a four-parameter binding model. Nearly 100% of CP3C virions, but only 40–50% of CR54 virions, can be neutralized by immune sera. Red dashed lines show how IC_{HM} is derived for animal 08D161. **c**, **d**, Neutralization curves of 9 viral variants using serum from one animal (**c**) or CD4-Ig (**d**). The parent virus into which mutations were made is listed, along with the amino acids at positions 23, 45, 47 and 70. **e**, All variants were assayed using serial dilutions of sera from all 40 Env-immunized animals. Shown is the maximum fraction of each virus that was neutralized (determined by regression analysis). Red letters indicate amino acid substitutions compared to the parent virus. The numbers above the graphic indicate the mean resistant fraction for each virus.

another clone from the E660 swarm, is a neutralization-resistant virus with IAK at these positions. Sera from immunized animals completely neutralized CP3C, with an inhibitory concentration potency (IC_{HM}, defined as the dilution giving half-maximal inhibition) that varied 1,000-fold (Fig. 3a). In contrast, the same sera only achieved a maximum of ~50% inhibition of CR54. Importantly, this shows that CR54 is a heterogeneous population of virions despite being genetically clonal: half of the virions are easily neutralized by antisera, and half are completely resistant.

We introduced variants of four amino acids into CP3C and CR54 Env to test which might be responsible for modulating neutralization resistance. These amino acid variations did not change the potency of the antisera (IC_{HM} varied less than twofold; Fig. 3c and Extended Data Fig. 5a), and did not change the sensitivity to neutralization by CD4-Ig (Fig. 3d). However, variant sequences affected the fraction of neutralization-resistant virions.

Despite the wide breadth of epitopes targeted by vaccine-elicited antisera (Extended Data Fig. 2), there was little variation in the fraction of each virus that was neutralization resistant (Fig. 3e). Moreover, an identical neutralization-resistant fraction was observed for a panel of SIV monoclonal antibodies directed near the CD4 binding site or the V1V2 loop (Extended Data Fig. 5b, c). Therefore, even clonally derived virions are remarkably heterogeneous: a fraction are easily neutralized, and the remainder are completely resistant to antibody-based neutralization.

Generation of this resistant Env phenotype was favoured by amino acid substitutions in the C1 region. By making point mutations, we showed that the T45A and R47K mutations individually result in increased resistance. Together, changing these two amino acids converts the sensitive CP3C Env to a nearly fully resistant phenotype, and the resistant CR54 to fully sensitive. For parsimony in subsequent analyses, we divided E660 viruses into two categories: viruses with both 45T and 47R ('TR'), which are putatively neutralization sensitive, and viruses with either 45A or 47K ('A/K'), which should be generally resistant to vaccine-elicited sera.

Deep sequencing and SGA of Env genes showed that ~20% of the E660 challenge swarm were neutralization-resistant A/K viruses (Fig. 2d). The same proportion was found among infecting T/F sequences in the control and Gag arms, demonstrating that there is no innate selection for or against A/K sequences. Furthermore, A/K infections resulted in the same peak and set point plasma VL, indicating that these viruses are no more or less fit than TR viruses (Extended Data Fig. 6). However, vaccine-elicited responses strongly selected against infection by TR viruses—such that, in the mac239 Env arm, the infrequent (neutralization-resistant) A/K variants comprised nearly 75% of T/F viruses.

We next computed the V_E against A/K and TR viruses separately. The TR (sensitive) variants are highly susceptible to vaccine-mediated

control, with a V_E of 90% (Fig. 2e). In contrast, the V_E against A/K viruses did not reach significance (Fig. 2f). Thus, the heterogeneous neutralization of even clonal SIV virions, programmed by C1 amino acid variations, represents a novel mechanism of immune escape from Env-specific antibodies.

Immune correlates of risk of infection

A panoply of cellular and humoral assays quantifying vaccine-elicited responses were performed at baseline, peak post-boost, and pre-challenge time points. We found strong associations between several antibody responses and probability of infection, but no associations between T-cell responses and delayed acquisition.

Given that the E660 swarm is comprised of both neutralization-sensitive (TR) and -resistant (A/K) genotypes, it made sense to analyse correlates in two ways: first, by including all infections, irrespective of variant; and second, by separating the two types of infections. Because the vaccine is largely ineffective against A/K viruses, pooling A/K-infected with TR-infected animals may mask potential correlates.

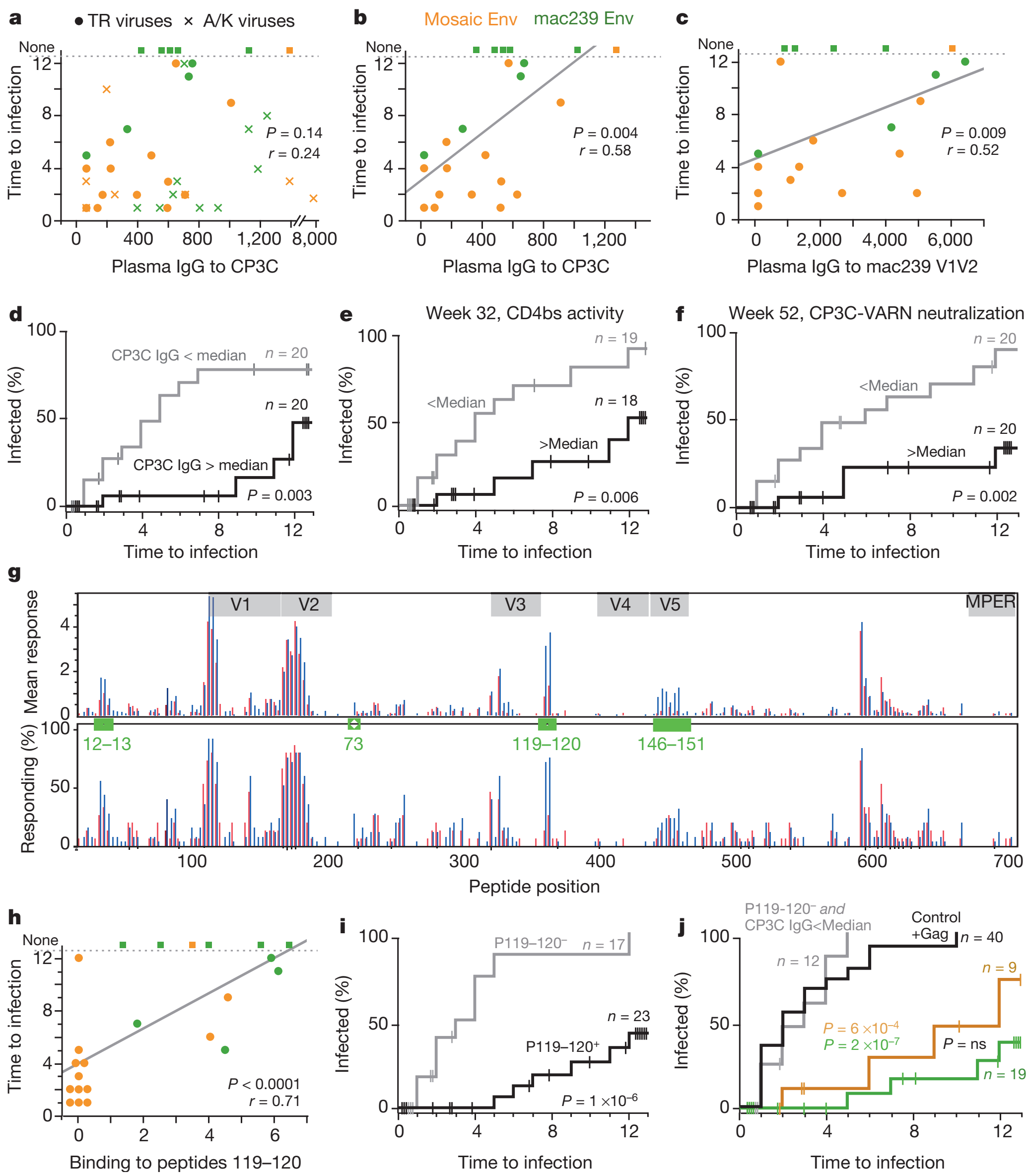


Figure 4 | Immunological correlates of risk. **a**, **b**, Week 52 plasma IgG against the CP3C envelope is graphed against time to infection (uninfected animals were assigned a value of 13). No significant correlation was found when all infection events were considered (**a**); however, by excluding animals infected solely with A/K viruses, a strong predictive relationship is seen (**b**). The line is from a linear regression for illustration; statistics are based on Spearman correlation. **c**, Week 52 plasma IgG against the mac239 V1V2 is significantly associated with protection against TR viruses, and also against all viruses (Extended Data Fig. 7b). **d**, Kaplan–Meier (KM) analysis was performed by dividing the 40 Env-immunized animals in two equal groups based on the anti-CP3C IgG responses (median = 570). Animals remaining uninfected or infected solely with A/K viruses were censored as shown by vertical lines. **e**, KM analysis comparing Env-immunized animals with higher vs lower week 32 serum activity against the CD4 binding site of envelope. **f**, KM analysis comparing Env-immunized animals with higher vs lower week 52 neutralization activity against virus pseudotyped with a CP3C Env containing a T45A mutation ('VARN'), a sequence shared by E543. **g**, The mean response (upper) and proportion of responders (lower) against each linear peptide is shown for animals grouped by time to infection: 1–3 challenges (red) vs 4 or more challenges (blue). Green boxes highlight regions potentially associated with protection identified by a Fisher's exact test; overlapping peptide numbers are in green, with sequences given in Supplementary Table 5. **h**, Average binding to the linear C3 peptides 119 and 120 (P119–120) correlates strongly with time to infection. **i**, KM analysis comparing Env-immunized animals with a positive response to C3 peptides to those with a negative response. **j**, KM analyses comparing all animals in the control and Gag arms (black), all Env-immunized animals having a CP3C IgG response below 570 and a negative C3 peptides response (grey), and animals in either Env arm having either antibody response.

The data shown in Fig. 4 illustrate these analyses. Among all 40 Env-vaccinated animals, plasma IgG binding to CP3C gp120 Env at the time of challenge did not correlate significantly with time to infection (Fig. 4a). In contrast, when we excluded animals who were infected solely with A/K viruses, we found a strong correlation with IgG binding to CP3C gp120 (Fig. 4b), but not other Envs (Extended Data Fig. 7a, b). We grouped all Env-immunized animals by those with an IgG response to CP3C above or below 570 (the median value, corresponding to an end point titre of approximately 1:1,000). Animals with the higher response had a 75% lower rate of infection by TR viruses (Fig. 4d).

Correlation with time to infection was also observed for plasma antibody avidity (Extended Data Fig. 7e), CD4-binding site activity (Fig. 4e) and neutralization of some viral strains (Fig. 4f). These data indicate that the quality of the antibody response is important. Thus, we investigated binding to specific regions within the Env.

By comparing peptide-binding data for animals grouped by time to infection (Fig. 4g), we identified four linear epitopes possibly associated with protection. There was a strong association between the breadth amongst these four epitopes and time to infection (Extended Data Fig. 8a–c). In contrast, there was no significant association with the breadth of response across all Env epitopes (Extended Data Fig. 8d). Thus, both breadth and magnitude of the response to selected epitopes are strong correlates of protection from acquisition.

The response to C3 (peptides 119+120) was the most significantly associated with protection, whether all viruses (Extended Data Fig. 8e, f) or just TR viruses were considered (Fig. 4h, i). This epitope corresponds to the α2 helix of Env and was identified as a neutralization target in HIV-1[20,21]. In a multivariable model, both IgG to CP3C ($P = 0.004$) and binding to the C3 peptides ($P = 0.02$) provided independent prediction of time to infection. We thus compared animals that had neither a response to the C3 peptides nor IgG to CP3C ($n = 12$, combining both Env arms) to animals with either response (mac239 Env: $n = 19/20$; mosaic Env: $n = 9/20$). For animals with neither antibody response, the rate of infection (12 infections in 27 exposures, 44%), and the proportion of infections with only A/K viruses (3/12, 25%) was not different from the control (unvaccinated) or Gag arms. In contrast, animals with either antibody response were primarily infected with resistant A/K viruses, and the V_E was >90% against TR viruses (Fig. 4j).

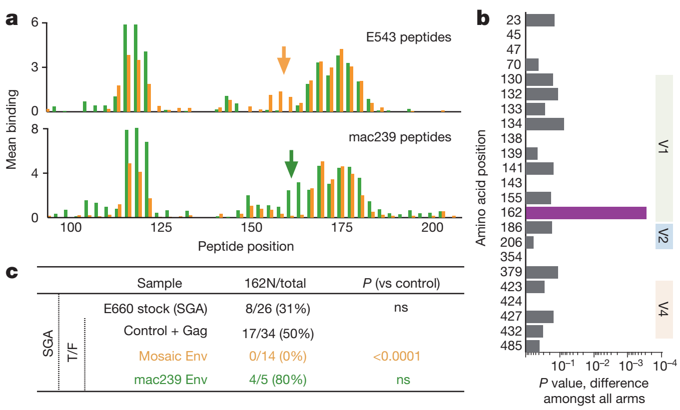


Figure 5 | Vaccine-mediated selection at V1V2. **a**, The binding of plasma from all 40 Env-immunized animals to linear 15-mer peptides spanning the V1V2 region of either E543 (top) or mac239 (bottom) was measured; bars represent the average binding for the 20 mosaic-immunized (orange) or the 20 mac239-immunized (green) animals. Arrows indicate an area of V1V2 showing vaccine-specific responses, encompassing amino acids 154–170. **b**, Sieving analysis was done as in Fig. 2c, but after excluding neutralization-resistant A/K viruses. The only significant association with immunization arm was at position 162. **c**, Representation of 162N (vs 162S) as determined by SGA for TR viruses in the swarm or in T/F. Note that all immunogens encode 162N, so selection is likely to be mediated against a neighbouring epitope; this epitope is found only in (one of) the mosaic immunogens, and occurs in linkage disequilibrium with 162N in the E660 swarm.

V1V2 and vaccine-specific sieving

In the human RV144 trial, antibody binding to HIV V1V2 was a primary (inverse) correlate of risk against infection. Similarly, antibody to the SIV V1V2 predicted protection against infection (Fig. 4c). The mosaic and mac239 immunogen sequences varied significantly in this region (Supplementary Table 1), and consequently elicited somewhat different antibody responses (Fig. 5a). To determine if these responses are associated with sieving, we analysed variation in T/F sequences (as in Fig. 2), after censoring vaccine-nonresponsive A/K viruses. This analysis revealed a strong selection associated with position 162 (Fig. 5b). The mosaic immunization completely selected against TR viruses with 162N (0/14 viruses, compared to 17/34 in the control; Fig. 5c). In contrast, the mac239 immunogen may have selected against the other variant (162S), although there were too few TR virus infections in this group to reach significance. Amongst A/K virus infections, there was no significant difference in representation of the 162N/S variants across vaccine arms. These data show that selection against V1V2 sequences by the SIV vaccines is limited to neutralization-sensitive viruses and, within those, selection is vaccine-sequence-specific and thus not broad.

Antibody escape mechanism in HIV

To assess whether our findings extend to HIV, we measured the inhibition of 51 distinct HIV-1 envelope pseudotyped viruses by the V1V2-specific monoclonal antibodies PG9 and PG16. As we saw for neutralization-resistant A/K SIV viruses, neutralization of some clonal HIV strains was incomplete; that is, a fraction of virions could not be neutralized (Fig. 6a). We examined the influence of sequence variation of these HIV envelopes on the fraction of neutralization-resistant virus (Fig. 6b); the most significant association was at position 47, with 47R viruses being sensitive (Fig. 6c). Sequence alignment with SIV envelope shows that position 47 in HIV is in a similar area of C1 as is position 47 in SIV (Fig. 6d); the similar signature (arginine vs lysine) indicates that a common mechanism of neutralization escape may be shared by SIV and HIV.

Discussion

Immune correlate studies that interrogate both virus sequences and immune responses can provide key insights on mechanisms of protection from HIV-1 acquisition. Using a nonhuman primate model

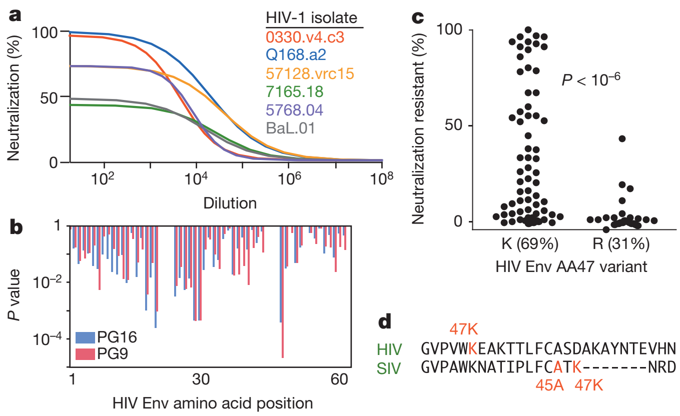


Figure 6 | C1 Sequences and HIV Env neutralization. **a**, Neutralization profiles of 51 different HIV-1 strains by the V1V2 antibodies PG9 and PG16 were determined. Example curves of PG16 on six viruses are shown. As for SIV A/K viruses, a variable fraction of each clonal virus is completely neutralization resistant; the remainder is highly sensitive. **b**, The influence of variants at each position in envelope on the fraction of neutralization resistant virus is shown as a P value from a Fisher's test; shown is the C1 region. **c**, The most significant association for all positions in Env was amino acid 47. The distribution of the fraction of neutralization-resistant virus is shown for the two variants, 47K and 47R. **d**, Alignment of SIV and HIV Env proteins in the middle of the C1 region, highlighting the positions of the neutralization signatures.

with a number of acquisition end points similar to large human efficacy studies, we demonstrated that an Env-elicited immune response is necessary and sufficient to provide protection from acquisition. We identified antibody-based correlates including responses to several epitopes. In our study, SIV Env T-cell mosaic immunogens elicited more effective T-cell responses, but less effective antibody responses. With respect to the virus, we identified a strong sieving effect of Env immunization, selecting for minor variants in the challenge swarm. And finally, we identify a sequence signature in the SIV Env, possibly shared by HIV, that programs the neutralization phenotype of the viruses through a mechanism affecting the entire antigenic surface of the protein.

Among our three vaccine groups, there was no association between protection from infection and protection from pathogenesis (for example, VL control). This suggests that humoral responses that effectively block acquisition are not necessarily correlated with cellular responses that control pathogenesis. Furthermore, we show that the Env-induced CTL suppressed acute viraemia better than Gag CTL, but suppressed chronic viraemia less effectively (Fig. 1b, d). Our data also show that vaccination resulted in reduced T/F viruses in breakthrough infections. This suggests that the primary mechanism of protection is by lowering the effective infectious dose, that is, *in vivo* neutralization.

Analysis of the sequences of breakthrough viruses revealed an amino acid signature, in the C1 region of Env, of viruses more likely to escape this neutralization. By creating point mutations that interconverted the neutralization profile of well-characterized viral envelopes, we defined a minimal two-amino-acid signature at positions 45 and 47 (TR vs A/K). Importantly, introduction of the A/K signature resulted in a fraction of clonally derived Env proteins having a 'global' antigenic change. This was manifested as resistance to polyclonal sera from dozens of animals, as well as resistance to monoclonal antibodies directed to the CD4 binding site or the V1V2 loops. Thus, the mechanism of resistance probably includes post-transcriptional modification, such as alternative glycosylation or folding, capable of masking the majority of epitopes on the viral Env.

We identified a hierarchy within this neutralization escape mechanism. This phenotype can occur for only a specific domain of the Env, such as for V1V2-directed antibodies against SIV (Extended Data Fig. 5b) and HIV (Fig. 6). This probably occurs through alternative glycosylation pathways restricted to this site. Resistance can also be global, affecting virtually any epitope, as we show for the SIV envelope (Extended Data Fig. 5a). A hierarchy was observed *in vivo*, in that sieving at the V1V2 domain was only observed in viruses lacking the global resistance phenotype (that is, in TR but not A/K viruses).

The observation that C1 amino acid variations can lead to alternative Env structures is consistent with data from ref. 22, where a single amino acid substitution was found to confer co-receptor dual tropism on mac239. Notably, the mutation responsible for the altered structure at the distant V3 loop was 47E—that is, within the signature we identified as conferring altered antigenicity upon SIV Env.

It is notable that all viruses in the mac251 swarm contain the resistant A/K signature. This may account for the weak correlation with vaccine-induced antibody in previous studies[11,14]. It is likely that the resistant Env form can be neutralized by antibodies targeting 'sites of vulnerability' (that is, rare epitopes conserved across all structures); for SIV, as it is in HIV[21], one of these may be the α2 helix. Antibody responses to this peptide were not only highly correlated with protection against TR viruses (Fig. 4g), but also showed a trend for protection against infection with the A/K viruses ($P = 0.07$). Likewise, the CD4-Ig molecule fully neutralized A/K viruses, suggesting that an appropriately targeted antibody to the CD4 binding site could have a similar effect.

By restricting our correlates analysis to exclude infections resulting from neutralization-resistant viruses (which are insensitive to the vaccine responses), we identified several strong correlates of risk of infection. All of these correlates derive from antibody measures, and include the magnitude of binding, the avidity of binding, and the breadth to selected epitopes of the SIV envelope. The importance of taking into account the virology is underscored by our analysis of the mosaic arm: despite this arm not achieving statistically significant protection overall (Fig. 1), we could identify active immune mechanisms (Fig. 4) as well as identify a mosaic immunogen-specific sieving effect in V1V2 (Fig. 5).

Our study provides insight into the possible reasons for the failure of HVTN505 and the limited protection in RV144[1]. Vaccination using our specific SIV Env expression vectors generated an antibody response ineffective against specific variants and protected against the subset of neutralization-sensitive viral variants (Fig. 2e). On the basis of data here, we propose that HVTN505 failed owing to an inability to elicit antisera that completely neutralized circulating HIV-1 strains, which are primarily neutralization-resistant. In contrast, the moderate success of RV144 suggests that antibodies were elicited that could neutralize some viruses circulating in that cohort; these sensitive viruses were susceptible to the vaccine-matched V1V2, leading to sieving. In any case, it will be critical to apply integrated analyses to HIV vaccine trials similar to what we did for this SIV study: that is, to clone and determine the neutralization profile of T/F viruses in the placebo arms (defining resistance of the circulating strains) and the active arms (to determine if the vaccine selectively blocked a subset of viruses), to optimally assess factors associated with vaccine-mediated protection.

Deciding which vaccine products to advance into large, expensive efficacy trials is difficult and complex. It is reasonable to postulate that any highly efficacious candidate will need to elicit antibodies targeting universal sites of vulnerability (that is, epitopes shared by the heterogeneous forms of even clonal virions), or to separately elicit antibodies targeting each structural form. Thus, understanding the biophysical basis for this viral heterogeneity will be crucial for designing vaccines capable of completely blocking HIV.

In conclusion, we identified a sequence signature of the SIV Env that distinguishes broadly neutralization-resistant viruses. By taking this signature of T/F viruses from breakthrough infections into account, we found several strong correlates of risk against infection, all based on antigen-specific antibody measurements—even for the mosaic vaccine arm that did not, upon initial analysis, reach statistically significant protection. We found that this signature, although probably not unique, is shared by SIV and HIV, and may underlie a fundamental mechanism of immune escape in both vaccinated and naturally infected subjects. Finally, our combined virological and immunological analyses provide insight into the biology of vaccine-mediated control, and lay a foundation for analysis and advancement of future HIV vaccines.

METHODS SUMMARY

Animals were handled in accordance with the standards of the American Association for the Accreditation of Laboratory Animal Care (AAALAC) and meet NIH standards as set forth in the Guidelines for Care and Use of Laboratory Animals. The animal protocol, VRC 10-332, was approved by the Vaccine Research Center IACUC. Functionality of all immunogens (mac239 and mosaic, Env and Gag) was confirmed by multiple assays. Animals were randomized into four groups of 20 based on *TRIM5α* alleles, gender, age and weight. Animals were challenged weekly with a dose of SIV_{smE660} previously shown to infect unvaccinated animals approximately 30% per exposure, as described[14]. Weekly challenges were initiated at week 53 (6 months after rAd5 boost), and were halted when an animal became PCR positive for viral RNA, or after 12 exposures. All immunological and virological assays performed for correlation analyses were qualified or validated, and performed by investigators blind as to group assignment and challenge outcome.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 September; accepted 21 November 2013.
Published online 18 December 2013.

1. Rerks-Ngarm, S. *et al.* Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N. Engl. J. Med.* **361,** 2209–2220 (2009).

2. Haynes, B. F. *et al.* Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *N. Engl. J. Med.* **366,** 1275–1286 (2012).
3. Tomaras, G. D. *et al.* Vaccine-induced plasma IgA specific for the C1 region of the HIV-1 envelope blocks binding and effector function of IgG. *Proc. Natl Acad. Sci. USA* **110,** 9019–9024 (2013).
4. Rolland, M. *et al.* Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. *Nature* **490,** 417–420 (2012).
5. Liao, H. X. *et al.* Vaccine induction of antibodies against a structurally heterogeneous site of immune pressure within HIV-1 envelope protein variable regions 1 and 2. *Immunity* **38,** 176–186 (2013).
6. Hammer, S. M. *et al.* Efficacy Trial of a DNA/rAd5 HIV-1 Preventive Vaccine. *N. Engl. J. Med.* **369,** 2083–2092 (2013).
7. Keele, B. F. *et al.* Low-dose rectal inoculation of rhesus macaques by SIVsmE660 or SIVmac251 recapitulates human mucosal infection by HIV-1. *J. Exp. Med.* **206,** 1117–1134 (2009).
8. Hidajat, R. *et al.* Correlation of vaccine-elicited systemic and mucosal nonneutralizing antibody activities with reduced acute viremia following intrarectal simian immunodeficiency virus SIVmac251 challenge of rhesus macaques. *J. Virol.* **83,** 791–801 (2009).
9. Lai, L. *et al.* Prevention of infection by a granulocyte-macrophage colony-stimulating factor co-expressing DNA/modified vaccinia Ankara simian immunodeficiency virus vaccine. *J. Infect. Dis.* **204,** 164–173 (2011).
10. Schell, J. B. *et al.* Significant protection against high-dose simian immunodeficiency virus challenge conferred by a new prime-boost vaccine regimen. *J. Virol.* **85,** 5764–5772 (2011).
11. Barouch, D. H. *et al.* Vaccine protection against acquisition of neutralization-resistant SIV challenges in rhesus monkeys. *Nature* **482,** 89–93 (2012).
12. Flatz, L. *et al.* Gene-based vaccination with a mismatched envelope protects against simian immunodeficiency virus infection in nonhuman primates. *J. Virol.* **86,** 7760–7770 (2012).
13. Lai, L. *et al.* SIVmac239 MVA vaccine with and without a DNA prime, similar prevention of infection by a repeated dose SIVsmE660 challenge despite different immune responses. *Vaccine* **30,** 1737–1745 (2012).
14. Letvin, N. L. *et al.* Immune and genetic correlates of vaccine protection against mucosal infection by SIV in monkeys. *Sci. Transl. Med.* **3,** 81ra36 (2011).
15. Lopker, M. *et al.* Heterogeneity in neutralization sensitivities of viruses comprising the simian immunodeficiency virus SIVsmE660 isolate and vaccine challenge stock. *J. Virol.* **87,** 5477–5492 (2013).
16. Fischer, W. *et al.* Distinct evolutionary pressures underlie diversity in simian immunodeficiency virus and human immunodeficiency virus lineages. *J. Virol.* **86,** 13217–13231 (2012).
17. Fischer, W. *et al.* Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nature Med.* **13,** 100–106 (2007).
18. Hudgens, M. G. & Gilbert, P. B. Assessing vaccine effects in repeated low-dose challenge experiments. *Biometrics* **65,** 1223–1232 (2009).
19. Hudgens, M. G. *et al.* Power to detect the effects of HIV vaccination in repeated low-dose challenge experiments. *J. Infect. Dis.* **200,** 609–613 (2009).
20. Gray, E. S. *et al.* Isolation of a monoclonal antibody that targets the alpha-2 helix of gp120 and represents the initial autologous neutralizing-antibody response in an HIV-1 subtype C-infected individual. *J. Virol.* **85,** 7719–7729 (2011).
21. Moore, P. L. *et al.* Limited neutralizing antibody specificities drive neutralization escape in early HIV-1 subtype C infection. *PLoS Pathog.* **5,** e1000598 (2009).
22. Del Prete, G. Q. *et al.* Derivation and characterization of a simian immunodeficiency virus SIVmac239 variant with tropism for CXCR4. *J. Virol.* **83,** 9911–9922 (2009).
23. Gray, R. T. A class of *K*-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Stat.* **16,** 1141–1154 (1988).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the following individuals: W. Shi, L. Wu, S.-Y. Ko, L. Wang and W.-P. Kong for immunogen construction; M. M. Donaldson, S.-F. Kao, D. Quinn, J. Owuor, K. Denison, H. Balachandran, C. Luedemann, W. T. Williams, G. Overman, A. Deal, C. Brinkley and L. Racz for technical assistance with immunology assays; A. Ault for assistance managing NHP studies; S. O'Connor for deep-sequencing data; M. Seaman for providing plasmids encoding E660 envelopes; F. McCutchan, J. Overbaugh and J. Kim for HIV-1 strains; and P. Gilbert for advice with using the Aalen and Johansen model. This work was supported by the Intramural Research Program of the Vaccine Research Center, NIAID, NIH; by NIH contracts HHSN261200800001E (B.F.K., W.G.) and HHSN27201100016C (D.C.M.); by NIH grant AI100645 (B.T.K., W.F.); and by the Bill and Melinda Gates Foundation grant OPP1032317.

Author Contributions M.R., R.A.S., R.A.K., G.J.N., N.L.L., S.S.R. and J.R.M. designed and supervised the study. W.F., Z.-Y.Y., B.T.K. and G.J.N. designed and manufactured immunogens. J.-P.M.T., N.L.L. and S.S.R. supervised nonhuman primate procedures. M.R., B.F.K., K.E.F., A.P.B., L.V.M., K.E.S., B.M.W., R.T.B., R.G., G.F., S.M.A., T.N.D., D.C.M., G.D.T., R.A.K. and J.R.M. supervised assays and performed primary data analysis. D.C.M. provided HIV-1 strains. S.D.S., R.D.M., H.C.W., C.L., M.K.L., L.V.M., L.S., K.E.S. and B.M.W. performed assays. M.R., W.G. and M.C.N. aggregated data and performed statistical analysis. M.R. and J.R.M. wrote the manuscript.

Author Information T/F sequences are deposited in GenBank under accession numbers KF602252–KF603880. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.R. (Roederer@NIH.gov).

METHODS

Animals. Animals were handled in accordance with the standards of the American Association for the Accreditation of Laboratory Animal Care (AAALAC) and meet NIH standards as set forth in the Guidelines for Care and Use of Laboratory Animals. The animal protocol, VRC 10-332, was approved by the Vaccine Research Center IACUC. All animals were Indian-origin rhesus macaques, male or female, approximately 3–5 years of age. Animals selected for the study were negative for MHC class I alleles Mamu-A01, -B08, and -B17[24]. Animals were typed by PCR for *TRIM5α* alleles, and categorized as having 0, 1 or 2 restrictive alleles[25]. 80 animals were randomized into four arms based on the following criteria: *TRIM5α* allele category, gender, age and weight. Blood was collected at regular intervals (weekly or biweekly). Peripheral blood mononuclear cells (PBMC) were prepared; a small number were reserved for phenotyping for absolute cell counts, and the remainder were viably cryopreserved (in fetal bovine serum containing 10% DMSO and stored in liquid phase nitrogen until analysis). Plasma was frozen at $-80\,^{\circ}$C for virological and serological analysis. Sample size ($n = 20$ per arm) was chosen to have an 80% probability to detect a vaccine efficacy of 50%[18].

Immunization. The design of the mosaic immunogens has been previously described[16]. Briefly, an input data set was assembled to include all available SIV_{sm} complete genome sequences that were either directly isolated from sooty mangabeys or (in a small number of cases) had been minimally passaged in tissue culture. So that viral challenges could reasonably approximate real-world post-vaccination exposure to unknown virus strains, we specifically excluded from the mosaic sequence input all sequences from the mac251 lineage (including mac239) as well as isolates of (and closely related sequences to) smE660. Mosaic sequence cocktails were generated sequentially, so that a single-sequence mosaic was generated first, and a second sequence was subsequently generated as the best complement to the first[26,27]. Coverage values of potential T-cell epitopes (as amino acid 9-mers) have been published[16]. Mosaic coding sequences were introduced into the same DNA and rAd5 backbones as the mac239 Env. All rAd5 vectors were produced by GenVec. The mosaic Env were given as gp160 in both the DNA prime and the rAd5 boost. The mac239 natural Env immunogens (DNA and rAd5) are identical to what was previously used[14], and was given as a gp140 in the DNA prime, and a gp145 in the rAd5 boost.

Functionality of all immunogens (mac239 and mosaic, Env and Gag) was confirmed by multiple assays. Expression of Env and Gag from DNA and rAd5 vectors in 293 cells in tissue culture was assessed by western blot analysis, and was found to be comparable. Immunogenicity of each vector/insert combination was confirmed in mouse studies.

Primates were immunized as previously described[14]: a total of four times; 4 mg of DNA was given intramuscularly at weeks 0, 4, and 8 and 10^{10} particles of rAd5 was given intramuscularly at week 28. The two Gag mosaic immunogens were mixed before administration, as were the two Env mosaic immunogens.

SIV challenge. Animals were challenged weekly with a dose of smE660 previously shown to infect unvaccinated animals approximately 30% per exposure, as previously described[14]. Weekly challenges were initiated at week 53 (6 months after rAd5 boost), and were halted when an animal became PCR positive for viral RNA, or after 12 exposures. There was no statistically significant change in the rate of infection within any group over time, indicating that infection was stochastic and there was no selection for innately resistant animals.

Assays. All immunological and virological assays performed for correlation analyses were performed by investigators blind as to group assignment and challenge outcome.

Virology. To quantify SIV viral load, viral RNA from plasma was isolated using a Qiagen QIAsymphony Virus/Bacteria Midi kit on Qiagen's automated sample preparation platform, the QIAsymphony SP. Viral RNA from 500 μl of plasma was eluted into 60 μl of elution buffer. All subsequent reactions were setup using Qiagen's automated PCR setup platform, the QIAgility. 25 μl of viral RNA was annealed to a target specific reverse primer 5′-CACTAGGTGTCTCTGCACTAT CTGTTTTG-3′ then reverse transcribed using SuperScript III RT (Invitrogen) and PCR nucleotides (Roche) along with RNase Out (Invitrogen) using an optimized version of the manufacturer's protocol. Resulting cDNA was treated with RNase H (Applied Biosystems) according to manufacturer's protocol. 10 μl of cDNA was then used to setup a real-time PCR using Gene Expression Mastermix (Applied Biosystems) along with target specific labelled probe 5′-/56-FAM/CT TCCTCAGTGTGTTTCACTTTCTCTTCTGCG/3BHQ_1/-3′ and forward 5′-GTC TGCGTCATCTGGTGCATTC-3′ and reverse primers 5′-CACTAGGTGTCTC TGCACTATCTGTTTTG-3′ (custom synthesis by Integrated DNA Technologies). Real-time PCR was performed on an Applied Biosystems StepOne Plus platform using the standard curve protocol. The RNA standard was transcribed from the pSP72 vector containing the first 731 bp of the SIV_{mac239} or SIV_{smE660} *gag* gene using the MEGAscript T7 kit (Ambion Inc.), quantitated by optical density (OD), and serially diluted to generate a standard curve. The quality of the RNA standard was assessed using an Agilent Bioanalyzer with RNA Nano 6000 chips (Agilent Inc.). The sensitivity of this assay has been shown to be 250 copies per ml.

The number of transmitted/founder (T/F) variants was determined by single genome amplification (SGA) of the full-length envelope gene as previously described[7]. The number of sequences analysed per animal was 21.2 ± 4.8 (mean ± s.d.), with a range of 10–38. There was no difference in number of sequences analysed by group.

All 1,629 sequences are deposited in GenBank under accession numbers KF602252–KF603880.

SIV envelope constructs. Sequences of the CP3C-P-A8 envelope (referred to in this paper as 'CP3C' for brevity) and CR54-PK-2A5 ('CR54') are shown in Supplementary Table 1. These sequences were used to produce protein for binding assays as well as pseudotyped viruses. Mutations were designed into each virus to create individual amino acid variants as listed in Fig. 3; the relevant portion of the envelope, encompassing the C1 region, is shown aligned in Supplementary Table 4. SIV Env mutant plasmids were generated by site-directed mutagenesis by GeneImmune Biotechnology.

Immunology. Intracellular cytokine staining for antigen-specific responses was performed using a qualified assay as described[28]. Cells were stimulated with overlapping 15mer peptides from Gag or Env from mac239 or smE543 (a clone similar to smE660). Data are shown for stimulations with E543 peptides. For breadth analysis, IFN-γ ELIspots were performed as described[29], using pools of 10 peptides from each protein.

Raw peptide microarray data (PepStar) were pre-processed and normalized as previously described[30]. Responses to peptides from mac239 or E543 were measured; data are shown for E543 only, except in Fig. 5a where data from both sets are shown. For each peptide, the mean binding from 10 control animals was subtracted from the value for each vaccinated animal. The distribution of resulting values was used to define a cut-off value of 1.2 for positivity: a large fraction of peptide responses constituted a near-normal distribution centred on 0 (after background subtraction); the 10th percentile of this distribution was -1.2; thus, $+1.2$ is an estimate of the 90th percentile of a completely negative response. For breadth analysis, positive responses to partially overlapping peptides were considered to comprise a single epitope.

SIV-specific humoral IgG and IgA levels were evaluated by a standardized antibody binding multiplex assay as previously described[31,32]. IgA levels were low and are shown only as MFI for the lowest dilution tested. IgG levels are shown as MFI AUC (area under the curve) computed over the dilutions in the linear range of the assay. Avidity was quantified by surface plasmon resonance (SPR) as previously described[2,33].

Viral neutralization assays. Neutralization was evaluated using three distinct assays. (1) Plasma neutralization of viral replication in PBMC was performed as previously described[14]. (2) Env-pseudovirus neutralization was measured using single-round-of-infection SIV Env-pseudoviruses with TZM-bl target cells stably expressing high levels of CD4 and the co-receptors CCR5 and CXCR4[34,35]. Tat-regulated luciferase gene expression was quantified to determine the reduction in virus infection. Neutralization curves were fit by nonlinear least squares regression, and the 50% inhibitory concentrations (IC_{HM}) was computed as the antibody concentration required to achieve 50% of maximal inhibition. (3) Replication competent SIV was used to infect TZM-bl cells as above, with cloned or uncloned swarm SIVs essentially as described[36]. Briefly, neutralization assays were performed with serial dilutions of heat-inactivated (56 °C, 1 h) samples. Diluted samples were pre-incubated with virus (~150,000 relative light unit equivalents) for 1 h at 37 °C before addition of cells. Following 48 h incubation, cells were lysed and luciferase activity determined using a microtitre plate luminometer and BriteLite Plus Reagent (Perkin Elmer). Neutralization titres are the reciprocal sample dilution or concentration (for sCD4) at which relative luminescence units (RLU) were reduced by 50% compared to RLU in virus control wells after subtraction of background RLU in cell control wells.

CD4 binding inhibition by sera was determined as described[37] with the following modifications. Plate-bound CP3C Env was incubated with or without a 1:5 dilution of plasma at 37 °C for 1 h. After washing, wells were incubated with 50 $\mu g\,ml^{-1}$ CD4-Ig–Biotin at 37 °C for 1 h. Plates were washed to remove excess CD4-Ig–Biotin and incubated with streptavidin horseradish peroxidase at 37 °C for 1 h. Inhibition was calculated as the fraction of the signal in wells with plasma to those without.

Statistics. The analyses presented here used a variety of techniques. Comparisons of continuous end points between groups were based on *t*-tests and analysis of variance, log-transformed when appropriate (for example, viral load). Comparisons of groups with respect to number of challenges until infection used the discrete time survival model assuming a leaky vaccine effect[18]. A comparison of the goodness-of-fit of possible models showed that the likelihood of the leaky model performed better than the all-or-none model (and the null hypothesis), and performed similarly to a

model that allowed both types of effects[18]. For the cumulative incidence of A/K vs TR viruses, we used nonparametric estimates that allowed for competing risks[23]. V_E against each virus type was computed by modifying the Hudgens and Gilbert leaky vaccine model[18] to account for two infection types, with P values computed from likelihood ratio tests. We did not formally test for heterogeneity of per-exposure probability. However, two pieces of evidence support lack of heterogeneity. First, the leaky model is a better fit than the all-or-none model (which should be sensitive to one type of heterogeneity). Second, we evaluated the risk of infection as a function of challenge number, and found no statistically significant change over time in any group.

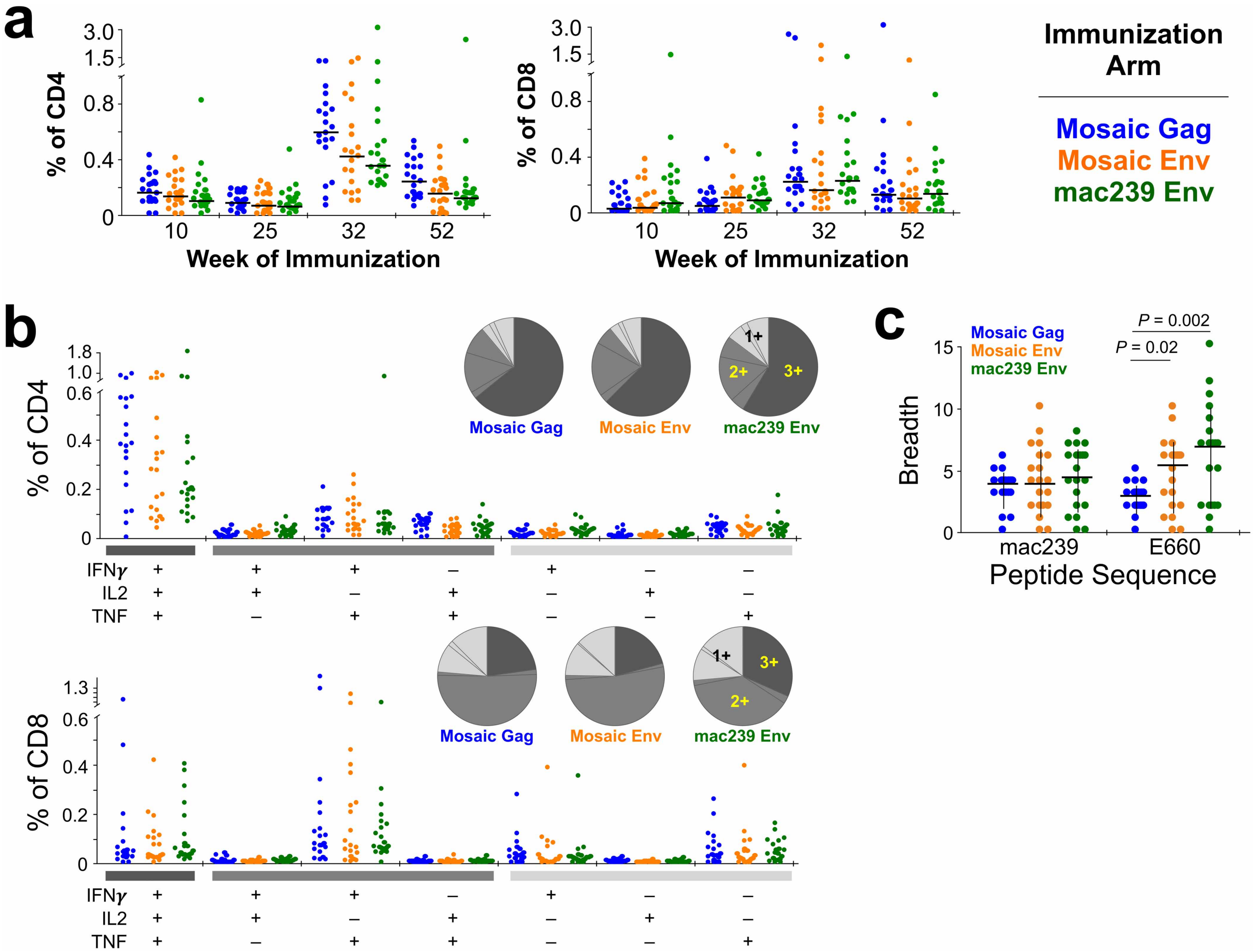
Associations between immunological measurements and number of challenges were based on similar models, with continuous predictors dichotomized at the median. Immunological predictors that showed some association using this method were investigated further using Cox Proportional Hazard models, both univariate and multivariate. All correlations were based on Spearman's rho (non-parametric), to handle variables with censored readings below or above assay limits, as well as to include uninfected animals in correlations based on number of infections. P values reported are not corrected for the number of comparisons between multiple immunological assays and outcome, except as noted.

In this paper we identified several potential correlates of risk of infection[38,39] that deserve further investigation and confirmation. Since these are not measureable in unvaccinated animals, the correlations with time to infection we report might be related to either differential effects of vaccines or to unspecified differences between the immune systems of the animals.

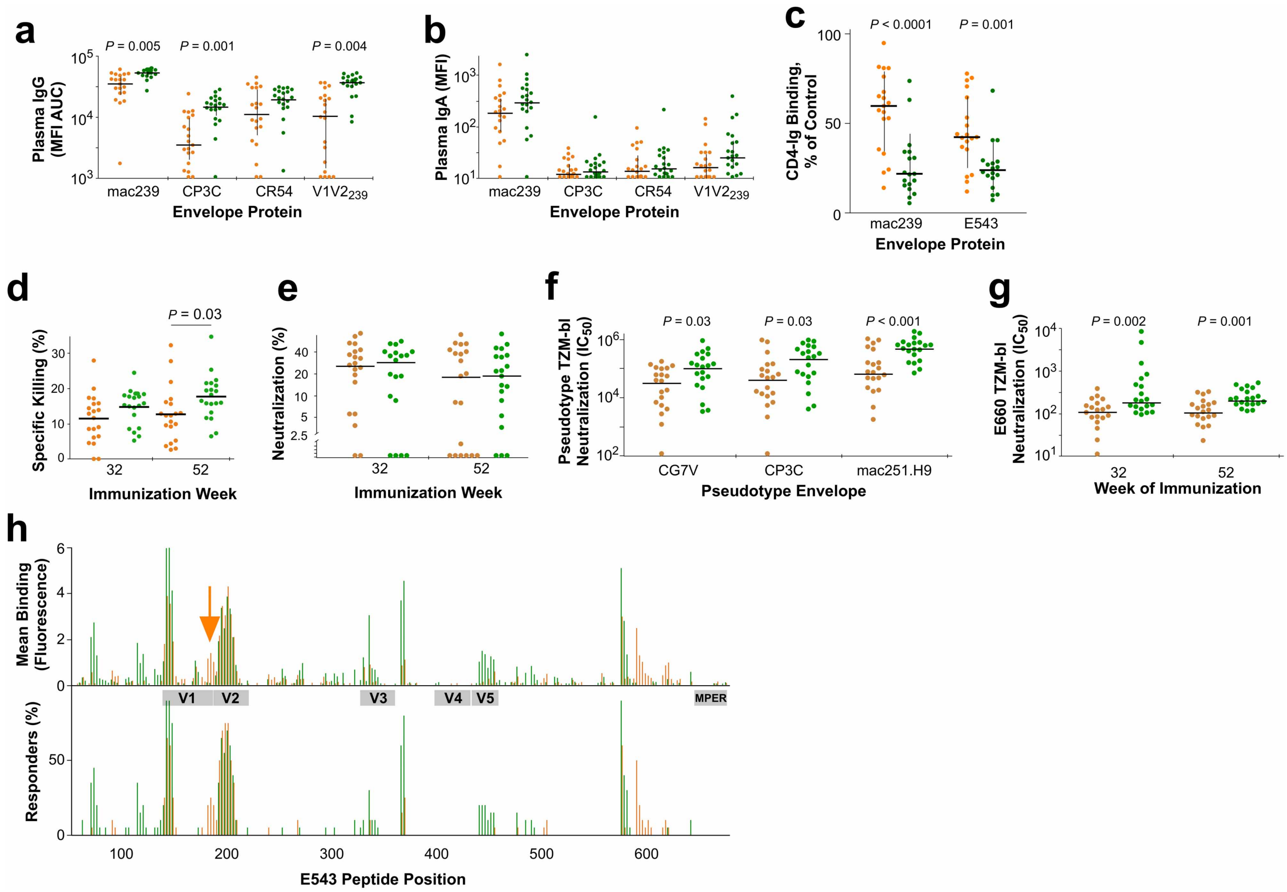
For sieve analysis, only Env positions with at least five variants amongst all sequences were considered. The distribution of variants at those positions passing this minimum threshold was compared between groups using a permutation-based version of the Fisher's exact test, using 10,000 permutations. P values reported in Fig. 2c and Extended Data Fig. 4 are not corrected for multiple comparisons, but positions 23, 45 and 47 (in both analyses) remain significant ($P \leq 0.002$) after such correction. Similarly, position 162 in Fig. 5b is significant after correction for multiple comparisons.

As expected from our stratification, *TRIM5α* alleles were found to have no effect on the conclusions of vaccine effects on protection; as follows: analysis of the discrete time-to-infection model using only *TRIM5α*-resistant animals did not change V_E. Cox proportional hazard modelling of time-to-infection by group did not change when *TRIM5α* was included as a covariate. Similarly, the importance of virus type infection (A/K vs TR) was unaffected by inclusion of *TRIM5α*. Finally, immunological correlates analyses (prediction of time-to-infection by antibody measures) did not change when *TRIM5α* was added as a covariate. The number of animals in each arm (out of 20) with homozygous *TRIM5α*-sensitive alleles was 8 (control), 9 (mosaic Env), 7 (mosaic Gag) and 8 (mosaic Gag).

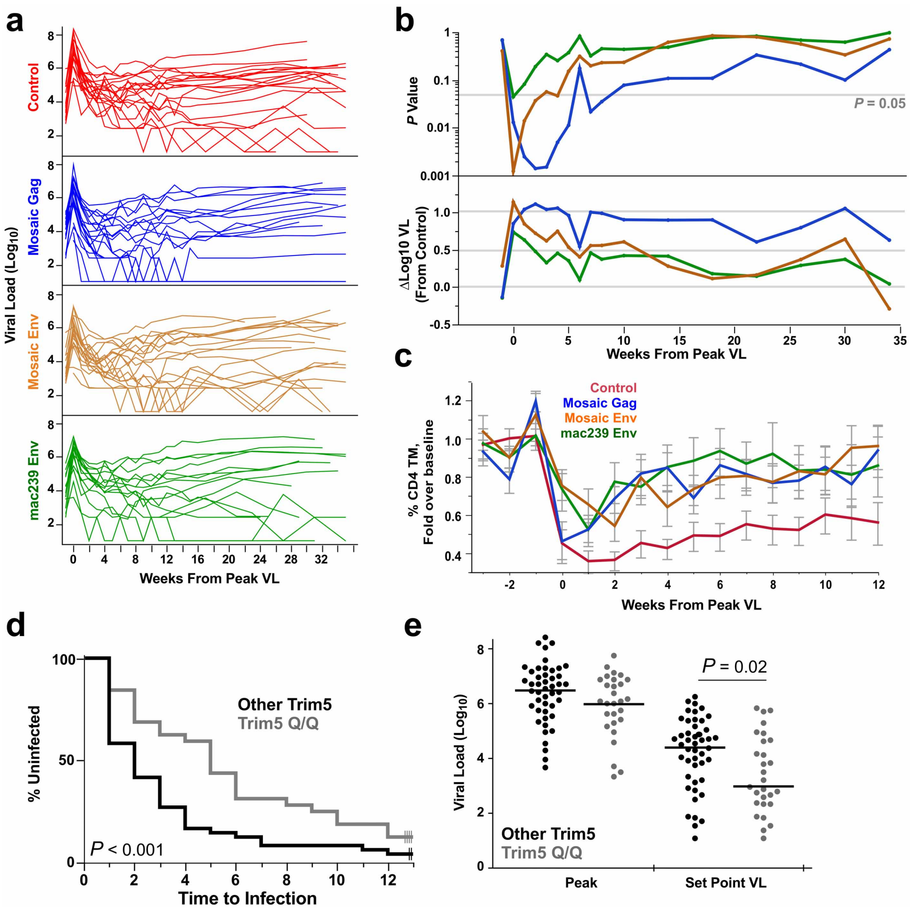
24. Lim, S. Y. *et al.* Contributions of *Mamu-A*01* status and *TRIM5* allele expression, but not *CCL3L* copy number variation, to the control of SIVmac251 replication in Indian-origin rhesus monkeys. *PLoS Genet.* **6,** e1000997 (2010).
25. Lim, S. Y. *et al.* TRIM5alpha modulates immunodeficiency virus control in rhesus monkeys. *PLoS Pathog.* **6,** e1000738 (2010).
26. Fenimore, P. W. *et al.* Designing and testing broadly-protective filoviral vaccines optimized for cytotoxic T-lymphocyte epitope coverage. *PLoS ONE* **7,** e44769 (2012).
27. Yusim, K. *et al.* Genotype 1 and global hepatitis C T-cell vaccines designed to optimize coverage of genetic diversity. *J. Gen. Virol.* **91,** 1194–1206 (2010).
28. Donaldson, M. M. *et al.* Optimization and qualification of an 8-color intracellular cytokine staining assay for quantifying T cell responses in rhesus macaques for pre-clinical vaccine studies. *J. Immunol. Methods* **386,** 10–21 (2012).
29. Santra, S. *et al.* Heterologous prime/boost immunizations of rhesus monkeys using chimpanzee adenovirus vectors. *Vaccine* **27,** 5837–5845 (2009).
30. Imholte, G. C. *et al.* A computational framework for the analysis of peptide microarray antibody binding data with application to HIV vaccine profiling. *J. Immunol. Methods* **395,** 1–13 (2013).
31. Bolton, D. L. *et al.* Comparison of systemic and mucosal vaccination: impact on intravenous and rectal SIV challenge. *Mucosal Immunol.* **5,** 41–52 (2012).
32. Tomaras, G. D. *et al.* Initial B-cell responses to transmitted human immunodeficiency virus type 1: virion-binding immunoglobulin M (IgM) and IgG antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. *J. Virol.* **82,** 12449–12463 (2008).
33. Alam, S. M. *et al.* Human immunodeficiency virus type 1 gp41 antibodies that mask membrane proximal region epitopes: antibody binding kinetics, induction, and potential for regulation in acute infection. *J. Virol.* **82,** 115–125 (2008).
34. Shu, Y. *et al.* Efficient protein boosting after plasmid DNA or recombinant adenovirus immunization with HIV-1 vaccine constructs. *Vaccine* **25,** 1398–1408 (2007).
35. Wu, L. *et al.* Enhanced exposure of the CD4-binding site to neutralizing antibodies by structural design of a membrane-anchored human immunodeficiency virus type 1 gp120 domain. *J. Virol.* **83,** 5077–5086 (2009).
36. Li, M. *et al.* Human immunodeficiency virus type 1 *env* clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *J. Virol.* **79,** 10108–10125 (2005).
37. Zhao, J. *et al.* Preclinical studies of human immunodeficiency virus/AIDS vaccines: inverse correlation between avidity of anti-Env antibodies and peak postchallenge viremia. *J. Virol.* **83,** 4102–4111 (2009).
38. Qin, L., Gilbert, P. B., Corey, L., McElrath, M. J. & Self, S. G. A framework for assessing immunological correlates of protection in vaccine trials. *J. Infect. Dis.* **196,** 1304–1312 (2007).
39. Plotkin, S. A. & Gilbert, P. B. Nomenclature for immune correlates of protection after vaccination. *Clin. Infect. Dis.* **54,** 1615–1617 (2012).



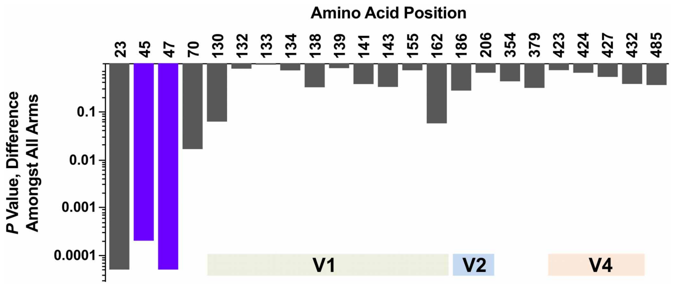
Extended Data Figure 1 | Cellular immunogenicity of vaccines. Gag- or Env-specific CD4 and CD8 T-cell responses measured by intracellular cytokine stimulation. Total T-cell responses were similar in all three active arms. **a**, Induction of T-cell responses are shown as the fraction of CD4 or CD8 memory T cells producing IFN-γ, IL2 or TNF in response to stimulation with overlapping peptides matched to the E660 challenge strain. Time points include peak post-DNA prime (week 10), pre-boost (week 25), peak post-rAd5 boost (week 32) and pre-challenge (week 52). **b**, The quality of the week 32 T-cell response is shown by the fraction of CD4 or CD8 cells responding to overlapping peptide pools matched to the E660 challenge swarm. There was no difference in the quality between any of the groups at any time point. **c**, Mosaic vaccination did not significantly improve the breadth of the T-cell response. Responses to pools of 10 overlapping peptides corresponding to mac239 Env or Gag, or the smE543 Env or Gag were tested for responses measured by ELISpot for the week 32 samples. Graphed is the number of positive pools (out of 23 for Env, and 13 for Gag) for each animal by group.



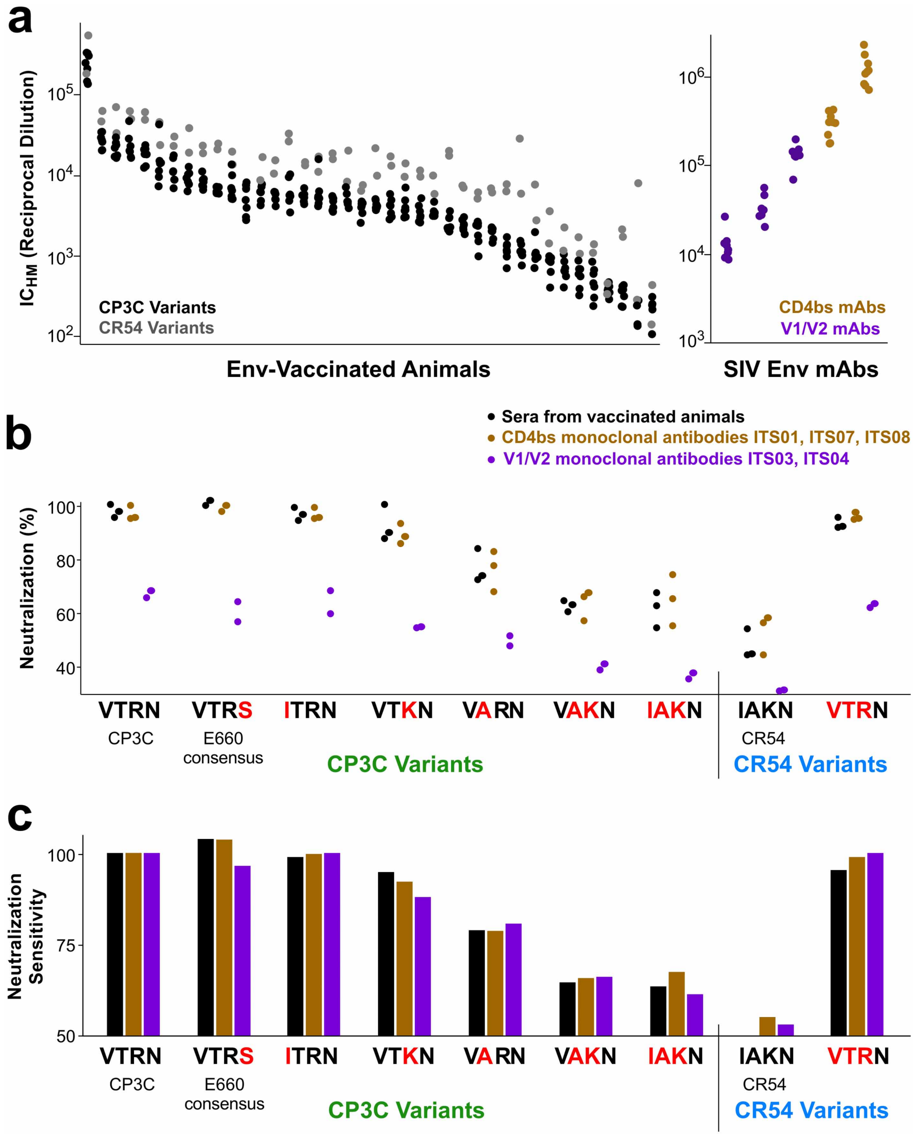
Extended Data Figure 2 | Humoral immunogenicity of vaccines. Mosaic immunization induced mildly lower humoral responses that were qualitatively different. **a**, **b**, Plasma IgG (**a**) or IgA (**b**) responses at week 32 were quantified against SIV envelope proteins derived from mac239, E660-CP3C, E660-CR54, or a mac239 V1V2 polypeptide expressed on a J08 scaffold. MFI, mean fluorescence intensity using a bead-based Luminex platform. AUC, area under the curve. **c**, CD4-binding site activity was measured by the ability of sera to cross-block CD4-Ig binding to mac239 or smE543 envelopes. **d**, Antibody-dependent cellular cytotoxicity mediated killing of SIV-infected target PBMC, shown as per cent specific killing. **e**, PBMC neutralization assay showing no substantial difference between immunization arms or time since vaccination. **f**, Neutralization by week 32 plasma was measured against three envelope-pseudotyped viruses. **g**, Neutralization of the E660 challenge stock using the TZM-bl indicator cell line. **h**, Week 32 plasma antibody binding to overlapping peptides spanning the SIV E543 envelope was quantified for the two envelope immunization arms. The mean response for all 20 animals in each arm (top) or the fraction of animals responding (bottom) is shown for peptides from the extracellular portion of the envelope. The arrow indicates an area near the V1V2 junction targeted by the mosaic but not the mac239 immunogen; several other areas, including C1, V3, C3 and V5, were better targeted by the mac239 immunogen.



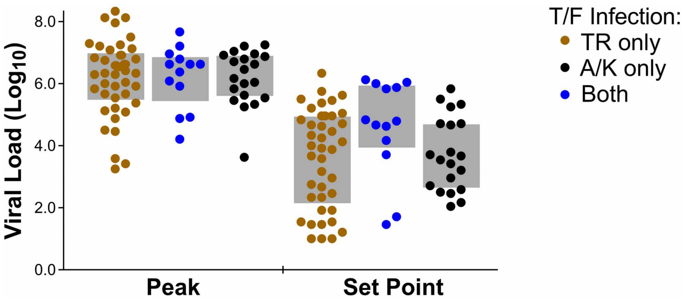
Extended Data Figure 3 | Viral pathogenesis and influence of *TRIM5α* alleles. **a**, Viral load (VL) was measured weekly until 12 weeks post peak and then monthly thereafter. Curves are shown for all 74 infected animals and are synchronized by the peak VL. **b**, For each time point, the distribution of VL in each immunization arm was compared to the control arm. The mean difference (lower) and significance of the difference (Student's *t*-test; upper) is graphed. **c**, The loss of CD4 cells following mucosal challenge is much more temperate than following intravenous challenge[31]. The most consistent measurable loss was for CD4 transitional memory cells ($CD45RA^{-}CCR7^{-}CD28^{+}$); the change in the frequency of these cells relative to the pre-infection average is shown. Other CD4 subsets showed less dramatic depletion. **d**, **e**, All 80 animals were grouped according to predicted resistance based on *TRIM5α* allelism (resistant: $TRIM5\alpha^{Q/Q}$; sensitive: all other combinations). A significant effect of genetics on acquisition (**d**) and pathogenesis (**e**) was observed. Animals were randomized equally into the four immunization arms based on *TRIM5α* genotype (for all homozygous and heterozygous genotypes).



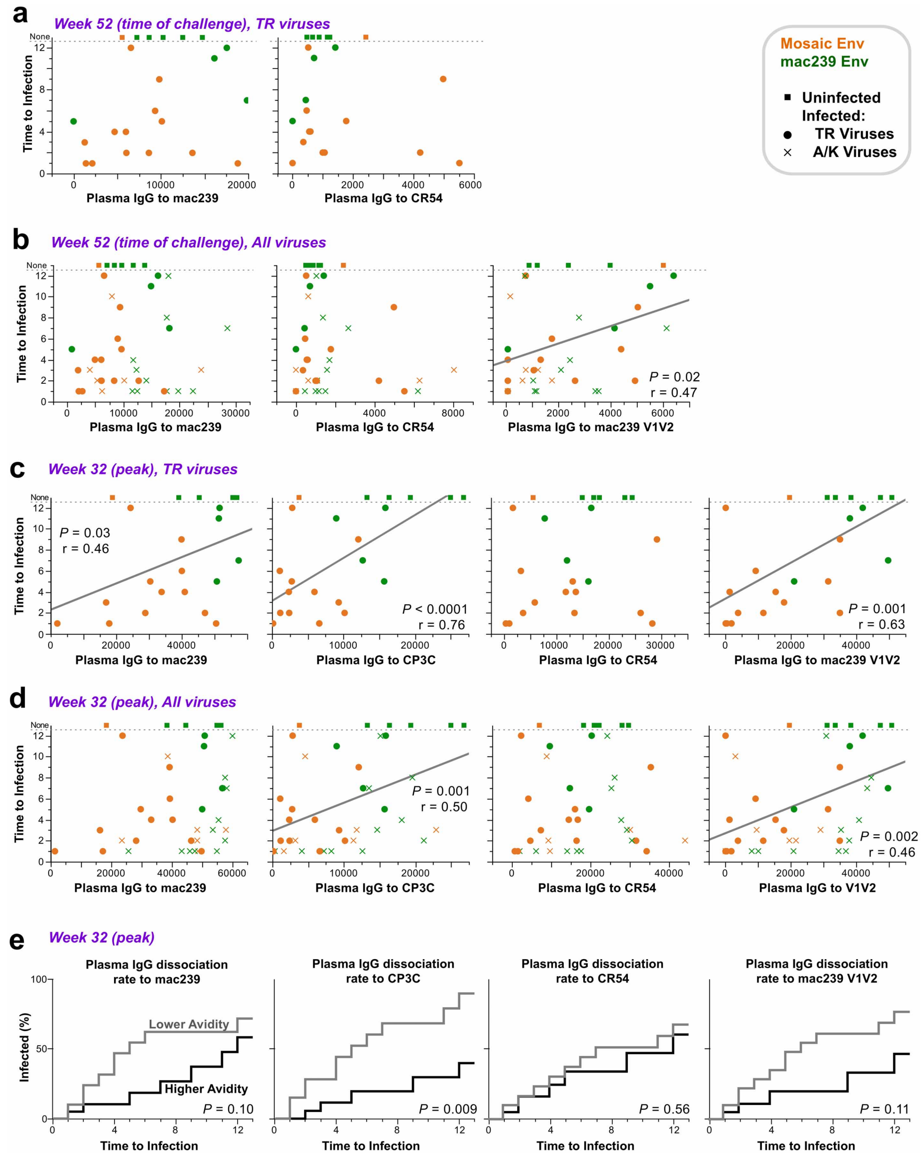
Extended Data Figure 4 | Transmitted/founder (T/F) selection in any vaccine arm. The number of T/F viruses with a variant from consensus was compared across all four arms for amino acid positions showing heterogeneity. A permutation test was used to compute the significance of a difference across all groups. The *P* values for positions 23, 45 and 47 remain significant after correction for multiple comparisons.



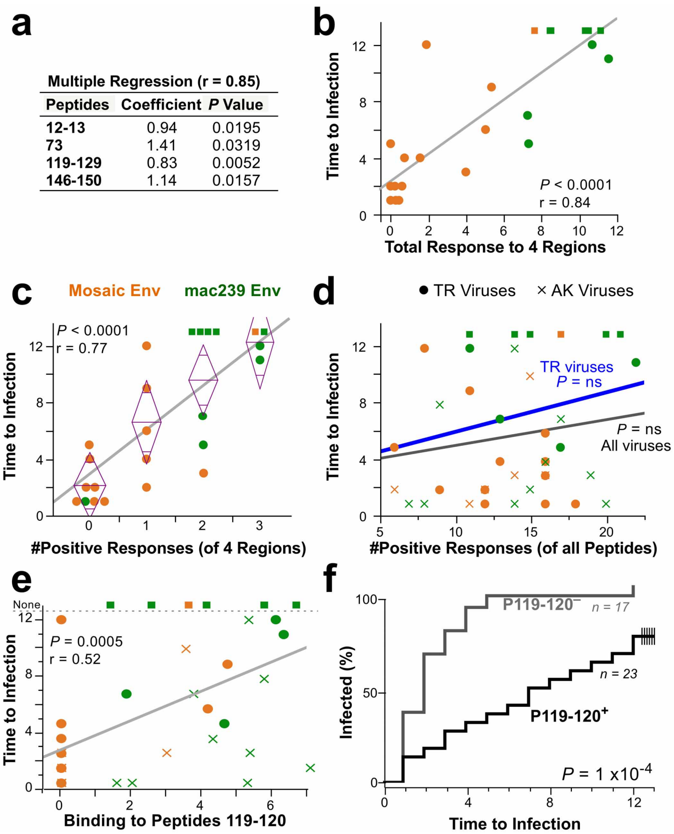
Extended Data Figure 5 | Neutralization sensitivity of variant envelopes. Nine envelope variants (Fig. 3) were evaluated for neutralization sensitivity by antisera from vaccinated animals (black or grey) and monoclonal antibodies to the CD4 binding site (brown) or the V1V2 loops (purple). **a**, The IC_{HM} (reciprocal concentration of antisera resulting in 50% of maximum neutralization) for all neutralization experiments is summarized by animal (left) or monoclonal antibody (right) for the seven CP3C variants and the two CR54 variants. The range of IC_{HM} across the viruses was less than twofold; that is, C1 sequence variations do not affect IC_{HM} but only the fraction of neutralization-resistant virions within each virus preparation (Fig. 3e). **b**, In a separate experiment, sera from three vaccinated animals and five monoclonal antibodies were compared. Note that the V1V2 antibodies only neutralize ~60% of the sensitive CP3C strain. **c**, Relative neutralization sensitivity was calculated by normalizing neutralization of each class of antibodies to 100% for CP3C.



Extended Data Figure 6 | Pathogenesis of TR and A/K viruses. Animals were divided into groups based on whether they were infected solely with TR viruses, A/K viruses, or both (that is, with multiple T/F per animal). Bars indicate the interquartile range of values. The peak and set point viral load did not differ according to which type of virus infected and replicated in the animal. In addition, no significant differences were observed when these data were split by vaccine arm.



Extended Data Figure 7 | Immunological correlates of risk, plasma IgG. **a**, **b**, Week 52 plasma IgG against the CP3C envelope is graphed against time to infection (uninfected animals were assigned a value of 13). Data are shown excluding A/K virus infections (**a**) or for all infections (**b**). Significant correlations are indicated by a linear least-squares regression line; statistics are nonparametric Spearman's tests. **c**, **d**, Similar analyses using week 32 (peak) plasma, for all TR infections (**c**), or all viral infections (**d**). **e**, Avidity to SIV envelopes was measured by Biacore; for each KM analysis, animals were divided in two equal groups based on having lower than median disassociation rate (high avidity) vs higher (low avidity), for TR infections.



Extended Data Figure 8 | Immunological correlates of risk, breadth of binding to linear peptides. **a**, A multivariable regression of time to infection vs responses to each of the four regions shown in Fig. 4g was performed. All four regions provided independent predictive power. **b**, The binding activity to all four regions was summed; the total response to these four epitopes showed a high correlation with time to infection. **c**, The number of the four regions with positive responses within each animal was computed (no animal responded to all four). The line indicates a linear regression; statistics are based on a nonparametric Spearman's test. **d**, The number of epitopes with positive responses across the entire envelope was computed for each animal. No correlation with protection (for all viruses or for only TR viruses) was seen with overall breadth. ns, $P > 0.05$. **e**, Average binding to the linear C3 peptides 119 and 120 correlates with time to infection for all animals, irrespective of virus. **f**, KM analysis comparing Env-immunized animals with a positive response to C3 peptides to those with a negative response.

ARTICLE

doi:10.1038/nature12940

Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection

Gilad Doitsh[1]*, Nicole L. K. Galloway[1]*, Xin Geng[1]*, Zhiyuan Yang[1], Kathryn M. Monroe[1], Orlando Zepeda[1], Peter W. Hunt[2], Hiroyu Hatano[2], Stefanie Sowinski[1], Isa Muñoz-Arias[1] & Warner C. Greene[1,2,3]

The pathway causing CD4 T-cell death in HIV-infected hosts remains poorly understood although apoptosis has been proposed as a key mechanism. We now show that caspase-3-mediated apoptosis accounts for the death of only a small fraction of CD4 T cells corresponding to those that are both activated and productively infected. The remaining over 95% of quiescent lymphoid CD4 T cells die by caspase-1-mediated pyroptosis triggered by abortive viral infection. Pyroptosis corresponds to an intensely inflammatory form of programmed cell death in which cytoplasmic contents and pro-inflammatory cytokines, including IL-1β, are released. This death pathway thus links the two signature events in HIV infection—CD4 T-cell depletion and chronic inflammation—and creates a pathogenic vicious cycle in which dying CD4 T cells release inflammatory signals that attract more cells to die. This cycle can be broken by caspase 1 inhibitors shown to be safe in humans, raising the possibility of a new class of 'anti-AIDS' therapeutics targeting the host rather than the virus.

The progressive loss of CD4 T cells in HIV-infected individuals lies at the root of AIDS. Despite more than three decades of study, the precise mechanism(s) underlying the demise of CD4 T cells during HIV infection remains poorly understood and has been highlighted as one of the key questions in HIV research[1]. In almost all cases, loss of CD4 T cells has been linked to apoptosis, both in *in vivo*[2–6] and *ex vivo*[5,7,8] studies. However, various features of apoptotic cell death including maturation of executioner caspase 3, DNA fragmentation and plasma membrane permeabilization are commonly shared with other programmed cell death pathways[9]. Importantly, most studies have focused on the death of productively infected cells circulating in peripheral blood[10]. Very little is known about the death of 'bystander' CD4 T cells in tissues that are refractory to productive HIV infection. However, these resting CD4 T lymphocytes represent the main cellular targets encountered by HIV in lymphoid tissues[11–13].

To investigate how CD4 T cells die during HIV infection, we took advantage of an *ex vivo* human lymphoid aggregate culture (HLAC) system formed with fresh human tonsil or spleen tissues[13]. HLACs can be infected with a small number of viral particles in the absence of artificial mitogens, allowing analysis of HIV cytopathicity in a natural and preserved lymphoid microenvironment[12]. Infection of these cultures with HIV-1 produces extensive loss of CD4 T cells, but over 95% of the dying cells are abortively infected with HIV, reflecting their non-permissive, quiescent state. The HIV life cycle is attenuated during the chain elongation phase of reverse transcription, giving rise to incomplete cytosolic viral DNA transcripts. Cell death is ultimately caused by a cellular innate immune response elicited by these cytosolic DNA intermediates[11]. This response is associated with production of type I interferon and activation of both caspase 3 and caspase 1. Caspase 3 activation leads to apoptosis without inflammation[14], whereas caspase 1 activation can trigger pyroptosis, a highly inflammatory form of programmed cell death in which dying cells release their cytoplasmic contents, including inflammatory cytokines, into the extracellular space[9,15]. The consequences of apoptosis versus pyroptosis may affect HIV pathogenesis by influencing the state of inflammation and immune activation, but their relative contribution to CD4 T-cell death in lymphoid tissues had remained unexplored.

Host permissivity determines the form of cell death

Previous reports have implicated caspase 3 activation and apoptosis in most instances of cell death caused by HIV-1 (refs 3, 7, 8). To explore the role of caspase 1 in dying HIV-infected CD4 T cells, HLACs formed with freshly dissected human tonsillar tissues were infected with a GFP reporter virus (NLENG1), prepared from the X4-tropic NL4-3 strain of HIV-1. This reporter produces fully replication-competent viruses. An internal ribosome entry site (IRES) upstream of the *nef* gene preserves Nef expression and supports long terminal repeat (LTR)-driven GFP expression[16], allowing simultaneous quantification of HIV-1 infection and caspase activation in CD4 T cells. NL4-3 was selected because tonsillar tissue contains a high percentage of CD4 T cells that express CXCR4 (90–100%). Consistent with our previous report[11], infection with HIV-1 produced extensive depletion of 'bystander' non-productively infected CD4 T cells. No more than 4% of the CD4 T cells were productively infected with HIV-1, but most of the remaining CD4 T cells underwent abortive infection and ultimately died after four days in culture (Fig. 1a).

To determine the distribution of active caspase 1 and caspase 3 in the dying CD4 T cells, we used fluorescently labelled inhibitor of caspases (FLICA) probes with sequences targeted by specific activated caspases[17]. Interestingly, a large fraction of non-productively infected CD4 T cells exhibited activation of caspase 1. Conversely, essentially no caspase 1 activity was detected in the productively infected cells (Fig. 1b). Caspase 3 activity was markedly less abundant, and mainly confined to the productively infected subset of cells (Fig. 1c). Treatment with efavirenz (a non-nucleoside reverse transcriptase inhibitor, NNRTI) or AMD3100 (an inhibitor of CXCR4-dependent HIV entry) prevented activation of both caspases. Infection with the primary, dual-tropic 89.6 HIV isolate[18] produced similar results (Extended Data Fig. 1). The two FLICA probes appeared to bind their respective caspases with reasonable specificity based on exclusive caspase 3 staining in cells treated with staurosporine, a protein kinase inhibitor known to induce apoptosis versus robust

[1]Gladstone Institute of Virology and Immunology, 1650 Owens Street, San Francisco, California 94158, USA. [2]Department of Medicine, University of California, San Francisco, 505 Parnassus Avenue, San Francisco, California 94143, USA. [3]Department of Microbiology and Immunology, University of California, San Francisco, 505 Parnassus Avenue, San Francisco, California 94143, USA.
*These authors contributed equally to this work.

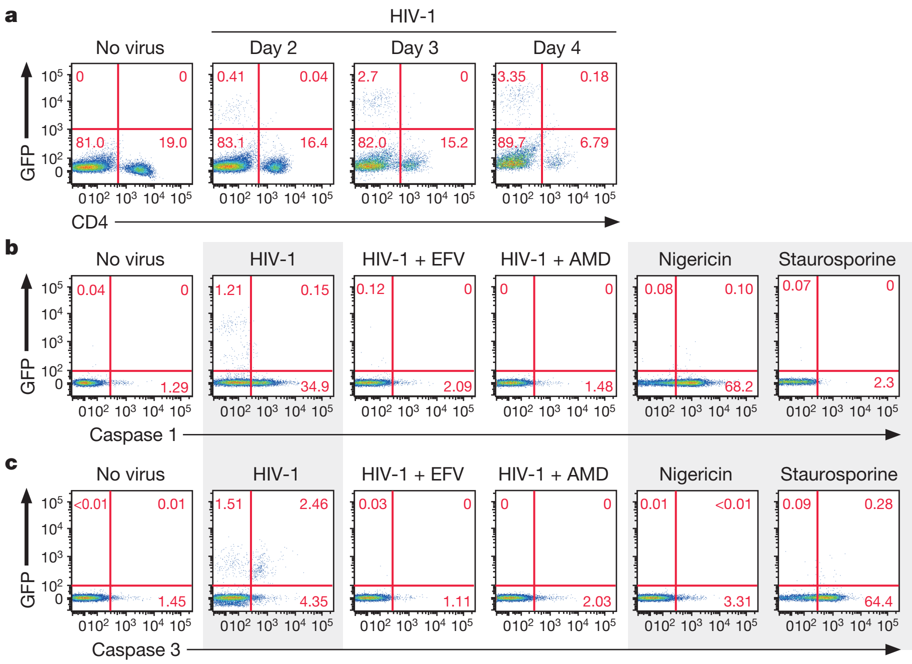


Figure 1 | Host permissivity determines the CD4 T-cell death pathway employed following HIV infection. **a**, Kinetics of spreading viral infection versus depletion of CD4 T cells after infection of HLACs with a replication-competent HIV reporter virus encoding GFP. The relative proportion of CD8 T cells was not altered (not shown). Consistent with our previous report, HIV-infected HLACs contain a small number of productively infected cells, whereas almost all of the dying cells are abortively infected[11]. **b**, Abortively infected CD4 T cells exclusively activate caspase 1. Nigericin induces abundant caspase 1 activation in uninfected cells. **c**, Productively infected CD4 T cells activate caspase 3 but not caspase 1. The samples (**b**, **c**) represent cells from the same infected tonsil culture. Efavirenz (EFV) and AMD3100 were added to the indicated cultures before HIV infection. These data are representative of four independent experiments performed with tonsil cells isolated from four different donors.

caspase 1 staining in cells treated with the cationic ionophore nigericin which promotes NLRP3 inflammasome assembly, caspase 1 activation and pyroptosis[19].

Healthy lymphoid CD4 T cells express pro-IL-1β

IL-1β activity is controlled at several levels including pro-IL-1β expression, processing and secretion. Pro-inflammatory stimuli induce expression of pro-IL-1β whereas processing and release are regulated by caspase 1 activation in inflammasomes[20]. The signals required for caspase 1 activation and release of IL-1β differ between immune cells. In circulating human blood monocytes, caspase 1 is constitutively active[21]. Stimulation of these cells with lipopolysaccharide (LPS) promotes pro-IL-1β expression leading to the rapid release of bioactive IL-1β. In contrast, macrophages and dendritic cells require a second signal to activate caspase 1 (ref. 22). Nigericin can function as this second signal activating caspase 1 in LPS-primed macrophages[19,23]. Surprisingly, nigericin alone proved sufficient to activate caspase 1 in uninfected lymphoid CD4 T cells (Fig. 1b) and to promote the release of the 17-kDa bioactive form of IL-1β (Fig. 2a). Treatment with monensin, a different monovalent cationic ionophore, or A23187, a calcium ionophore, did not promote mature IL-1β release[23,24]. Maturation and secretion of the bioactive form of IL-1β was inhibited by Z-VAD-FMK (a pan-caspase inhibitor), Z-WEHD-FMK or Z-YVAD-FMK (two independent caspase 1 inhibitors, which also block other inflammatory caspases—caspase 4 and caspase 5), but not by Z-FA-FMK (a negative control for caspase inhibitors) indicating that caspase 1 activation was required.

Pro-IL-1β expression in human tonsil and spleen HLACs was examined next. Western blotting analysis revealed large amounts of intracellular pro-IL-1β in both untreated tonsil and spleen HLACs (Fig. 2b). Removal of dead cells by Ficoll-Hypaque density centrifugation resulted in an even higher intracellular pro-IL-1β signal, indicating that these normal lymphoid tissues constitutively express high levels of pro-IL-1β. The presence of pro-IL-1β in spleen indicated that expression in tonsil is not solely caused by infection (tonsillitis). Fractionation of the lymphocytes present in these HLACs revealed high levels of intracellular pro-IL-1β in isolated CD4 T cells, but not in CD8 T-cell or B-cell populations.

Most tonsillar CD4 T cells express CXCR4, but only around 5% of these cells also express CCR5 (refs 12, 25). When CCR5-positive and CCR5-negative lymphoid CD4 T-cell subsets were isolated and studied, the CCR5-expressing cells displayed much higher levels of intracellular pro-IL-1β (Fig. 2b). The CCR5-expressing CD4 T cells also released notably more 17 kDa IL-1β into the supernatant after infection with HIV-1 (Fig. 2c). These results suggest that most of the mature form of IL-1β is released by the small population of CCR5-expressing CD4 T cells. The resident CCR5-expressing cells in lymphoid tissues are primarily memory CD4 T cells, which might be more permissive for productive HIV infection[26]. However, the activation status of these cells varied (Fig. 2d). Two-thirds exhibited a memory phenotype as determined by surface expression of CD45RO, but only a small fraction of these cells were permissive to productive infection with either X4-tropic or R5-tropic HIV-1 strains (Extended Data Fig. 2). Notably, lymphoid CCR5-expressing CD4 T cells also express CXCR4 and thus can be targeted by either X4 or R5-tropic HIV-1 strains[12,27,28]. Memory T cells continually recirculate within lymphoid tissues scanning for presentation of their cognate antigen[29–31]. It seems likely that many of these cells have returned to a sufficient state of quiescence that they are susceptible to abortive HIV infection and thus could contribute importantly to chronic inflammation through the release of bioactive IL-1β .

CD4 T-cell death by HIV-1 is mediated by pyroptosis

Caspase 1 is a pro-inflammatory caspase whose catalytic activity is tightly regulated by signal-dependent auto-activation within inflammasomes[20]. Inflammasome-dependent caspase 1 activity results in a highly inflammatory form of cell death known as pyroptosis, primarily described in myeloid cells infected with intracellular bacterial pathogens[9,15,32]. Pyroptosis is caspase 1 dependent by definition and occurs independently of other pro-apoptotic caspases[9,32]. Based on our finding that caspase 1 is activated in lymphoid CD4 T cells following abortive HIV infection, we investigated whether pyroptosis is triggered within these cells.

Fresh HLACs were infected with HIV-1 and cultured for 12 h to initiate viral spread and then treated with various caspase inhibitors or controls. Extensive and selective depletion of CD4 T cells occurred in untreated, HIV-infected cultures after 3 days. However, treatment with either pan-caspase or caspase 1 inhibitors prevented the depletion of CD4 T cells as efficiently as the viral inhibitors efavirenz and AMD3100 (Fig. 3a). Inhibitors of caspase 3 or caspase 6 and the control compound did not prevent CD4 T-cell depletion. Necrostatin-1, a RIP1 inhibitor, did not inhibit CD4 T-cell depletion (Extended Data Fig. 3a, b), indicating that cell death does not reflect necroptosis. Analysis of spleen cells yielded similar results (Extended Data Fig. 3c). Inhibiting type-I interferon signalling with neutralizing antibodies directed against IFNα/β

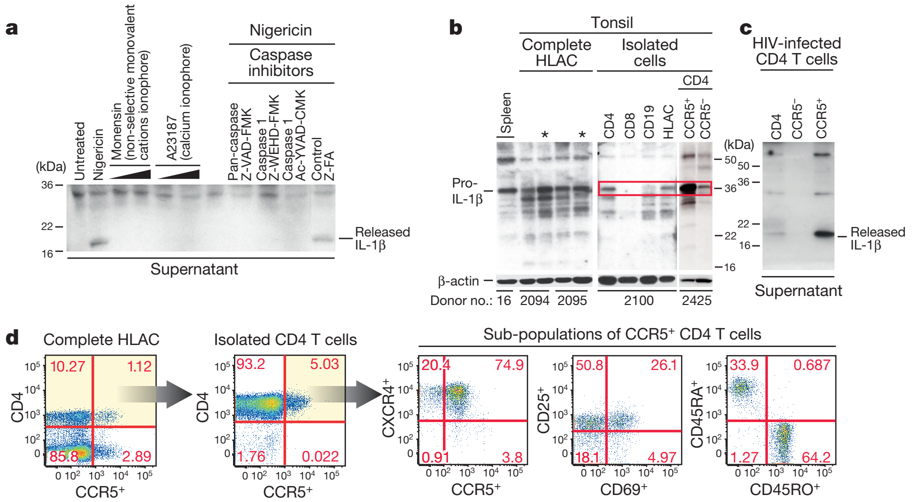


Figure 2 | Lymphoid CD4 T cells are primed to mount an inflammatory response and constitutively express high levels of pro-IL-1β. **a**, A secondary inflammatory stimulus by nigericin induces lymphoid CD4 T cells to process and release bioactive IL-1β. Supernatants from cell cultures were filtered to remove all remaining cells and subjected to SDS–polyacrylamide gel electrophoresis (SDS–PAGE) immunoblotting analyses for bioactive 17-kDa IL-1β. **b**, High levels of constitutive pro-IL-1β are selectively expressed in lymphoid CD4 T cells. Levels of intracellular pro-IL-1β were assessed in HLACs from fresh tonsils or spleen tissue from different donors. Asterisks indicate samples in which dead cells were removed. CD4 T cells were positively isolated from HLAC. Cells were lysed and analysed for pro-IL-1β expression. **c**, Nearly all bioactive IL-1β produced by HIV-infected lymphoid CD4 T cells is released from CCR5-expressing cells. Indicated CD4 T-cell populations were isolated from HLAC and infected with HIV-1. Supernatants of cultures were filtered and analysed for bioactive 17-kDa IL-1β. **d**, HLACs were characterized for expression of memory and activation markers by flow cytometry. The majority of CCR5-expressing CD4 T lymphocytes exhibit a memory phenotype. All CCR5-expressing CD4 T cells co-express the CXCR4 receptor.

receptor did not prevent CD4 T-cell death (Extended Data Fig. 4), indicating that this antiviral response is not critical for the innate-immune-mediated onset of programmed cell death. Distinct from apoptosis, pyroptosis features cellular swelling, plasma membrane rupture and release of intracellular content into the extracellular milieu[15], including cytosolic enzymes like lactate dehydrogenase (LDH)[33]. LDH release was readily detected after HIV infection (Fig. 3b), and was blocked by two antiviral inhibitors, efavirenz and AMD3100 and by a caspase 1 inhibitor, but not by a caspase 3 inhibitor. Thus, the form of cell death associated with abortive HIV infection appears to involve caspase 1 activation and the release of cytoplasmic contents. Caspase 1 inhibitors also prevented death of CCR5-expressing CD4 T cells in HLACs infected with a CCR5-dependent strain of HIV-1 (Fig. 3c). Inhibition of cell death by the caspase 1 inhibitor was as effective as the CCR5 receptor antagonist TAK779, suggesting that most CCR5-expressing CD4 T cells in the culture are dying by caspase-1-mediated pyroptosis. These findings are consistent with the large amounts of bioactive IL-1β released by these cells after HIV-1 infection.

Because caspase inhibitors are not exquisitely specific, we designed short hairpin (shRNA) vectors to silence the expression of caspase 1, the ASC (PYCARD) adaptor, which recruits pro-caspase 1 to inflammasome complexes[20], caspase 3 and NLRP3 (Extended Data Fig. 5). For these experiments, a third generation shRNA-encoding lentiviral vector (shRNA LV) pSico[34], bearing an EF1α:mCherry reporter expression cassette was used. To relieve the resistance of lymphoid CD4 T cells to shRNA LV infection, target cells were initially challenged with lentiviral particles harbouring Vpx (Vpx-VLPs), which induce proteasomal degradation of SAMHD1 in non-permissive human resting CD4 T cells[35]. Infections with Vpx-VLPs did not lead to activation of resting CD4 T cells, as measured by surface expression of the CD69 and CD25 activation markers (not shown). The shRNA LV particles and Vpx-VLPs were pseudotyped with a CXCR4-tropic Env of HIV-1, which supports efficient fusion to quiescent CD4 T lymphocytes[36]. Under these conditions, infection with shRNA LVs markedly suppressed expression of a variety of targeted genes whereas the scrambled shRNA LV control did not (Fig. 3d). We next investigated whether any of these shRNA LVs inhibited pyroptosis induced by nigericin. Nigericin induced massive pyroptosis in mCherry positive CD4 T cells infected with scrambled and caspase 3 shRNA LV particles, but this response was blocked by the caspase 1, ASC or NLRP3 shRNA LV particles (Fig. 3e). Next, the effect of these shRNAs on CD4 T-cell death elicited by HIV-1 was examined. HIV-1 infection caused extensive death of mCherry-positive CD4 T cells expressing shRNAs against scramble, caspase 3 and NLRP3, but not caspase 1 or ASC. Thus, cell death occurring during abortive HIV infection appears to be mediated through caspase 1 dependent pyroptosis involving an inflammasome that contains ASC but lacks NLRP3.

HIV-1 stimulates caspase 1 to secrete IL-1β

To independently confirm that abortive HIV-1 infection leads to the activation of caspase 1, we investigated the appearance of the active p10 subunit of caspase 1. As controls for pyroptosis and apoptosis, uninfected cells were treated with either nigericin or staurosporine, respectively. An active 10 kDa subunit of caspase 1 (p10) was detected in the lysates of HIV-infected cultures as well as in nigericin-treated cells, and in blood monocytes in which caspase 1 is constitutively active[21]. Treatments with viral or caspase 1, but not caspase 3, inhibitors prevented caspase 1 cleavage (Fig. 3f). These findings confirm the induction of caspase 1 in quiescent CD4 T cells following abortive infection with HIV-1. Caspase 3 activation in these infected cultures was markedly less abundant (Extended Data Fig. 6). To test whether caspase 1 activation leads to proteolytic maturation of pro-IL-1β, we used various caspase inhibitors and analysed the culture media for the mature 17-kDa form of IL-1β. Interestingly, release of mature IL-1β was completely inhibited by a pan-caspase inhibitor and by two different caspase 1 inhibitors (Fig. 3g). Inhibitors of apoptotic caspases, caspase 3, caspase 6 or caspase 8, or necrostatin did not interrupt this release. Similar findings were observed using a quantitative IL-1β enzyme-linked immunosorbent assay (ELISA) (Extended Data Fig. 7a). Thus, caspase 1 activation is specifically required for the release of bioactive IL-1β in lymphoid CD4 T cells infected with HIV-1. In accord with shRNA analyses, treatment with four separate NLRP3 inhibitors did not prevent release of bioactive IL-1β by HIV-1 (Fig. 3g), nor CD4 T-cell death by HIV-1 (Extended Data Fig. 7b, c).

In vivo evidence for HIV-mediated pyroptosis

To extend our *ex vivo* HLAC findings, we next examined fresh lymph node tissue obtained from a consenting untreated subject infected with

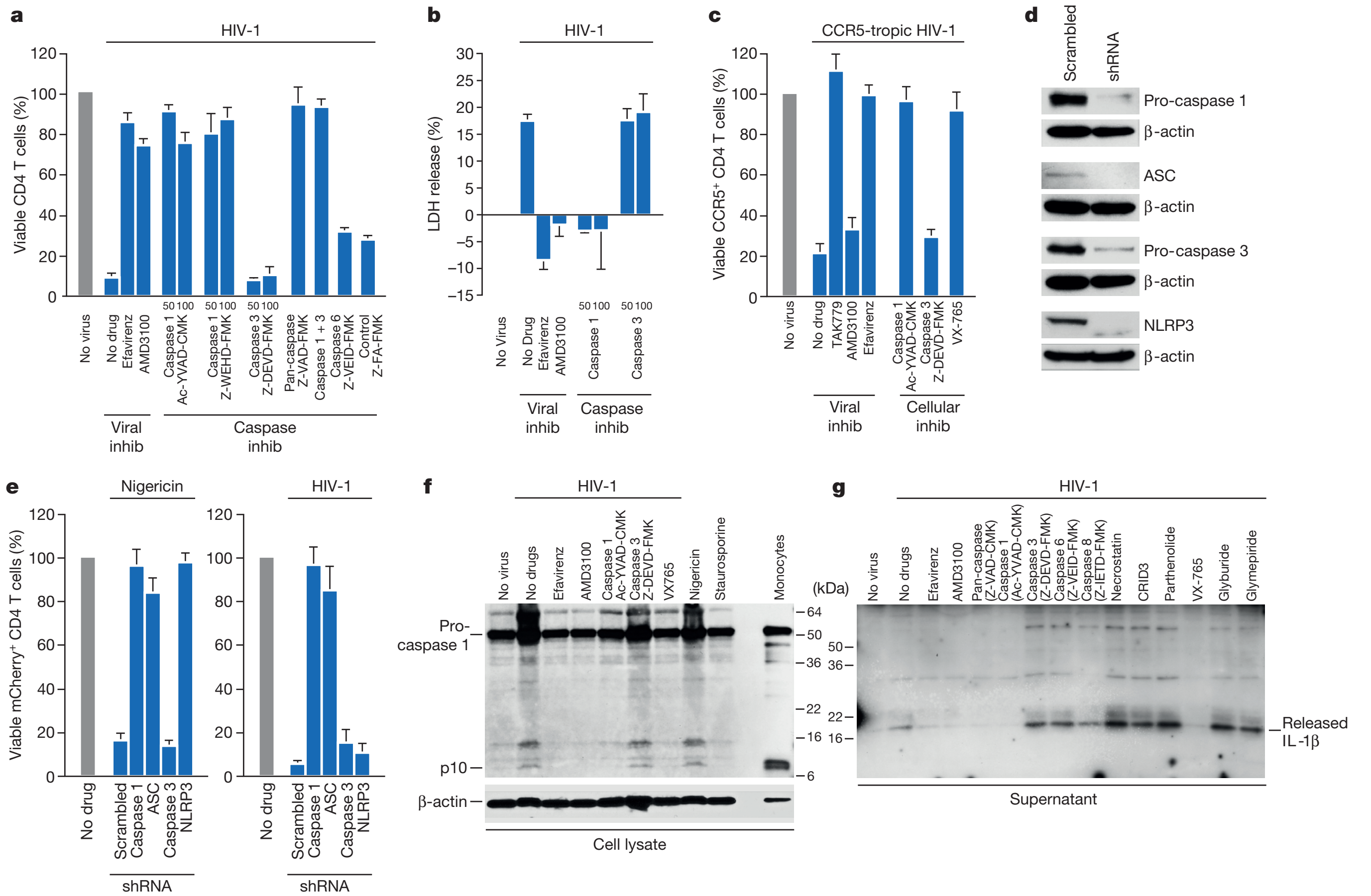


Figure 3 | Death of HIV-infected lymphoid CD4 T cells and release of bioactive IL-1β are controlled by caspase 1. **a**, **b**, Caspase 1 inhibitors are sufficient to prevent CD4 T-cell death in HIV-infected HLACs. Viable CD4 T cells were counted by flow cytometry and supernatants were analysed for levels of cytoplasmic LDH enzyme release[33]. **c**, Infection with CCR5-dependent HIV-1 induces pyroptosis of lymphoid CD4 T cells. Death of CCR5-expressing CD4 T cells is prevented by caspase 1 inhibitors and TAK779, but not by the CXCR4 antagonist, AMD3100. Due to the small number of target CCR5-expressing cells, this experiment was performed by overlaying tonsil cells on a monolayer of 293T cells that had been transfected with an R5-tropic proviral HIV-1 clone, as previously described[11]. The co-culture conditions for the R5 virus experiment induced no activation of the overlaid cells. **d**, Efficient repression of target genes by shRNA-coding lentiviral vectors. **e**, shRNA LV designed to silence either caspase 1 or ASC, key components of the pyroptotic pathway, protect lymphoid CD4 T cells from death by nigericin treatment or HIV-1 infection. To specifically assess non-productively infected cells, cultures were treated with 3′-azido-3′-deoxythymidine (AZT) before infections with HIV-1. **f**, Caspase 1 cleavage in HIV-infected CD4 T cells is blocked by specific caspase 1 inhibitors. **g**, Inhibitors of caspase 1, but not NLRP3, prevent release of bioactive IL-1β from HIV-infected lymphoid CD4 T cells. Error bars represent s.e.m from at least three independent experiments using tonsil cells from at least three different donors. Protein analyses represent results from three independent experiments using tonsillar CD4 T cells from three different donors.

R5-tropic HIV and displaying a high viral load and a low CD4 T-cell count. *In situ* immunostaining revealed a distinct zone of HIV p24 Gag expression between the mantle zone and germinal centres, where activated CD4 T and B cells proliferate (Ki-67) and interact in the follicles (Fig. 4). Conversely, staining for caspase 1 revealed abundant activity in the surrounding paracortical zone (CD3) comprised primarily of resting CD4 T cells. Staining of uninfected tonsil or spleen (not shown) tissues revealed no such positive signals (Extended Data Fig. 8). Because this antibody reacts with both the active p20 component of caspase 1 and pro-caspase 1, we cannot completely exclude the possibility that abortive HIV-1 infection produced a localized increase in pro-caspase 1 expression. However, large amounts of IL-1β were also detected in the paracortical zone, particularly in the extracellular space between the T cells, as well as the cell death marker annexin V. In sharp contrast, active caspase 3 staining was limited to the areas in the germinal centre where HIV-1 p24 Gag expression was detected. These findings strongly agree with the HLAC results (Fig. 1b) indicating that caspase 3 activity occurs in a set of productively infected cells, anatomically separated from most of the resting CD4 T cells undergoing abortive infection, caspase 1 activation, IL-1β processing and pyroptosis.

A clinically safe drug blocks pyroptosis by HIV-1

Identifying pyroptosis as the predominant mechanism mediating CD4 T-cell depletion during HIV infection provides novel targets, such as caspase 1, for potential therapeutic intervention. The role of caspase 1 in the chronic inflammatory response has attracted therapeutic interest[37]. VX-765 is a caspase 1 inhibitor that has been tested in chronic epilepsy and psoriasis (Extended Data Fig. 9a)[38–41], and found in a phase IIa trial (http://clinicaltrials.gov/ct2/show/NCT01048255) to be safe and well tolerated over the six-week length of the trial[42]. In our studies, VX-765 inhibited IL-1β secretion by nigericin-induced lymphoid CD4 T cells (Extended Data Fig. 7b), indicating it efficiently blocks caspase 1 activity in these cells. VX-765 also blocked caspase 1 cleavage (Fig. 3f), IL-1β secretion (Fig. 3g) and CD4 T-cell death in HIV-infected tonsillar and splenic HLACs (Figs 3c and 5a, b and Extended Data Fig. 9b). Cell death was not markedly inhibited by VRT-043198 (the active form of the VX765 pro-drug), probably because of reduced cellular permeability[38]. HIV-1 infection was not restored to productive infection when caspase 1 was blocked (Extended Data Fig. 10). These findings demonstrate that a small-molecule inhibitor of caspase 1, shown to be safe in humans, suppresses CD4 T-cell death and inflammation elicited in lymphoid tissues by HIV-1.

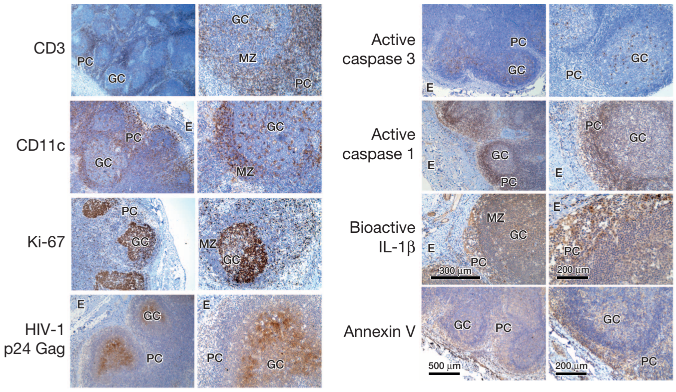


Figure 4 | Distinct regions of caspase 1 and caspase 3 activity in lymph node of a patient chronically infected with R5-tropic HIV. Inguinal lymph node was collected from a 50-year-old immunosuppressed HIV-1 infected subject during the chronic phase of disease. The patient was first identified with HIV in 1985, has not been on anti-retroviral therapy and displayed a CD4 count of 156 cells per μl and viral load of 85,756 copies per ml at the time of lymph node resection (see also Extended Data Fig. 8). E, epithelium; GC, germinal centre; MZ, mantle zone; PC, paracortical zone.

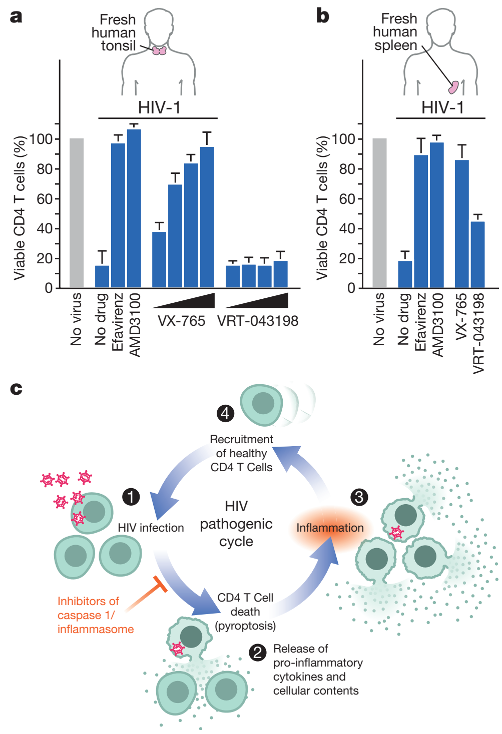


Figure 5 | Targeting caspase 1 via an orally bioavailable and safe drug prevents lymphoid CD4 T-cell death by HIV-1. a, **b**, VX-765 efficiently blocks CD4 T-cell death in HIV-infected tonsillar and splenic lymphoid tissues. No toxicity was observed at any of these drug concentrations. Error bars represent s.e.m. from three independent experiments using tonsil or spleen cells from three different donors. **c**, Pyroptosis in HIV-infected lymphoid tissues may establish a chronic cycle of CD4 T-cell death and inflammation, which attracts new CD4 T cells and ultimately contributes to disease progression and tissue damage. Inhibitors of caspase 1 such as VX-765 may inhibit pyroptosis in a manner that both preserves CD4 T cells and reduces inflammation.

Discussion

HIV's lethal attack on its principal cellular target, the CD4 T cell, has been generally attributed to apoptosis[2-8,43]. We now demonstrate that the permissivity status of the host cell dictates the pathway through which lymphoid CD4 T cells die following HIV infection. Specifically, when HIV infects permissive, activated CD4 T cells, cell death occurs silently through caspase-3-dependent apoptosis. Conversely, when either R5- or X4-tropic HIV abortively infects non-permissive, quiescent CD4 T cells from lymphoid tissue, these cells die by caspase-1-dependent pyroptosis, an intensely inflammatory form of programmed cell death. Our recent studies have identified IFI16 as the host DNA sensor that recognizes the incomplete HIV reverse transcripts thereby initiating activation of caspase 1 (ref. 44). In most human lymphoid tissues including tonsil, lymph node and spleen, the activated and permissive subset of cells represents 5% or less of the total CD4 T cells, whereas the non-permissive quiescent cells represent 95% or more of the targets encountered by HIV[12,25]. Thus, in sharp contrast to previous studies[2-8,10], caspase-1-mediated pyroptosis, not caspase-3-mediated apoptosis, appears predominantly responsible for driving CD4 T-cell death following HIV infection of these lymphoid tissues. These findings are further supported by analysis of fresh lymph nodes from subjects infected with R5-tropic HIV, in which caspase 1 and IL-1β are detected in the paracortical zone that is rich in resting CD4 T cells, whereas caspase 3 activity is detected in the anatomically distinct germinal centres where productively infected cells are found.

Our studies also highlight how lymphoid CD4 T cells are selectively primed to mount inflammatory responses as evidenced by constitutive expression of cytoplasmic pro-IL-1β. This is particularly prominent within the CCR5-expressing subset of lymphoid CD4 T cells. The pyroptotic death of these cells would lead to high level release of IL-1β potentially further fuelling chronic inflammation.

Pyroptosis probably promotes the rapid clearance of various bacterial infections by removing intracellular replication niches and enhancing the host's defensive responses through the release of pro-inflammatory cytokines and endogenous danger signals. However, in pathogenic chronic inflammation, such as in HIV infection, pyroptosis is not a protective response and does not lead to clearance of the primary infection. In fact, pyroptosis appears to create a pathogenic vicious cycle in which dying CD4 T cells release inflammatory signals that attract more cells into the infected lymphoid tissue to die and to produce more inflammation[45] (Fig. 5c). These events establish a chronic state of inflammation that probably fuels disease progression and tissue injury[46]. Chronic inflammation might also promote maintenance of the latent HIV reservoir through the dysregulated action of the IL-7 or IL-15 cytokines stimulating homeostatic proliferation of memory CD4 T cells. In this regard, it will be interesting to assess to what extent pyroptosis persists in lymphoid tissues of HIV-infected subjects on effective anti-retroviral therapy.

The depletion of CD4 T cells and the development of chronic inflammation are signature processes in HIV pathogenesis that propel disease progression[47]. Our studies now reveal how pyroptosis provides an unexpected link between these two disease-promoting processes. In non-pathogenic infections in which simian immunodeficiency virus (SIV) infects its natural non-human primate hosts, caspase 3 apoptosis in productively infected cells may signal for most of the cell death rather than caspase 1, thus reducing inflammation. The pathogenic cycle of cell death and inflammation created by pyroptosis obligately requires the activation of caspase 1. As such, it may be possible to break this pathogenic cycle with safe and effective caspase 1 inhibitors. These agents could form a new and exciting 'anti-AIDS' therapy for HIV-infected subjects in which the treatment targets the host instead of the virus.

METHODS SUMMARY

Human tonsil or splenic tissues were obtained from the National Disease Research Interchange and the Cooperative Human Tissue Network and processed as previously described[11]. Dead cells within the complete HLACs were first removed by Ficoll-Hypaque gradient centrifugation. CD4 T cells ($CD3^+$) were isolated from HLACs by positive selection using CD4 microbeads (Miltenyi) as described[11]. CCR5-expressing CD4 T cells were positively separated (PlusCellect R&D Systems), from CD4 T cells negatively isolated from complete HLACs (STEMCELL Technologies, EasySep Human $CD4^+$ T-cell Enrichment Kit). In shRNA experiments, infections with R5-tropic HIV-1, and when splenic cells that are extremely refractory to HIV-1 infection were used, we modified the infection system by overlaying HLAC cells on a monolayer of 293T cells that had been transfected with HIV-1 proviral clones, as previously described[11]. Flow cytometry data were collected on a FACS Calibur (BD

Biosciences) and analysed with FlowJo software (Treestar). HIV-1 viruses were generated by transfection of proviral DNA into 293T cells using calcium phosphate.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 August; accepted 5 December 2013.

Published online 19 December 2013.

1. Thomas, C. Roadblocks in HIV research: five questions. *Nature Med.* **15,** 855–859 (2009).
2. Muro-Cacho, C. A., Pantaleo, G. & Fauci, A. S. Analysis of apoptosis in lymph nodes of HIV-infected persons. Intensity of apoptosis correlates with the general state of activation of the lymphoid tissue and not with stage of disease or viral burden. *J. Immunol.* **154,** 5555–5566 (1995).
3. Finkel, T. H. *et al.* Apoptosis occurs predominantly in bystander cells and not in productively infected cells of HIV- and SIV-infected lymph nodes. *Nature Med.* **1,** 129–134 (1995).
4. Huang, M. B., James, C. O., Powell, M. D. & Bond, V. C. Apoptotic peptides derived from HIV-1 Nef induce lymphocyte depletion in mice. *Ethnic. Dis.* **18,** S2–30–37 (2008).
5. Røsok, B. I. *et al.* Correlates of apoptosis of CD4$^+$ and CD8$^+$ T cells in tonsillar tissue in HIV type 1 infection. *AIDS Res. Hum. Retroviruses* **14,** 1635–1643 (1998).
6. Gougeon, M. L. *et al.* Programmed cell death in peripheral lymphocytes from HIV-infected persons: increased susceptibility to apoptosis of CD4 and CD8 T cells correlates with lymphocyte activation and with disease progression. *J. Immunol.* **156,** 3509–3520 (1996).
7. Jekle, A. *et al. In vivo* evolution of human immunodeficiency virus type 1 toward increased pathogenicity through CXCR4-mediated killing of uninfected CD4 T cells. *J. Virol.* **77,** 5846–5854 (2003).
8. Grivel, J. C., Malkevitch, N. & Margolis, L. Human immunodeficiency virus type 1 induces apoptosis in CD4$^+$ but not in CD8$^+$ T cells in *ex vivo*-infected human lymphoid tissue. *J. Virol.* **74,** 8077–8084 (2000).
9. Lamkanfi, M. & Dixit, V. M. Manipulation of host cell death pathways during microbial infections. *Cell Host Microbe* **8,** 44–54 (2010).
10. Cooper, A. *et al.* HIV-1 causes CD4 cell death through DNA-dependent protein kinase during viral integration. *Nature* **498,** 376–379 (2013).
11. Doitsh, G. *et al.* Abortive HIV infection mediates CD4 T cell depletion and inflammation in human lymphoid tissue. *Cell* **143,** 789–801 (2010).
12. Eckstein, D. A. *et al.* HIV-1 actively replicates in naive CD4$^+$ T cells residing within human lymphoid tissues. *Immunity* **15,** 671–682 (2001).
13. Glushakova, S., Baibakov, B., Margolis, L. B. & Zimmerberg, J. Infection of human tonsil histocultures: a model for HIV pathogenesis. *Nature Med.* **1,** 1320–1322 (1995).
14. Ren, Y. & Savill, J. Apoptosis: the importance of being eaten. *Cell Death Differ.* **5,** 563–568 (1998).
15. Fink, S. L. & Cookson, B. T. Apoptosis, pyroptosis, and necrosis: mechanistic description of dead and dying eukaryotic cells. *Infect. Immun.* **73,** 1907–1916 (2005).
16. Levy, D. N., Aldrovandi, G. M., Kutsch, O. & Shaw, G. M. Dynamics of HIV-1 recombination in its natural target cells. *Proc. Natl Acad. Sci. USA* **101,** 4204–4209 (2004).
17. Bedner, E., Smolewski, P., Amstad, P. & Darzynkiewicz, Z. Activation of caspases measured in situ by binding of fluorochrome-labeled inhibitors of caspases (FLICA): correlation with DNA fragmentation. *Exp. Cell Res.* **259,** 308–313 (2000).
18. Collman, R. *et al.* An infectious molecular clone of an unusual macrophage-tropic and highly cytopathic strain of human immunodeficiency virus type 1. *J. Virol.* **66,** 7517–7521 (1992).
19. Mariathasan, S. *et al.* Cryopyrin activates the inflammasome in response to toxins and ATP. *Nature* **440,** 228–232 (2006).
20. Schroder, K. & Tschopp, J. The inflammasomes. *Cell* **140,** 821–832 (2010).
21. Netea, M. G. *et al.* Differential requirement for the activation of the inflammasome for processing and release of IL-1β in monocytes and macrophages. *Blood* **113,** 2324–2335 (2009).
22. Laliberte, R. E., Eggler, J. & Gabel, C. A. ATP treatment of human monocytes promotes caspase-1 maturation and externalization. *J. Biol. Chem.* **274,** 36944–36951 (1999).
23. Perregaux, D. & Gabel, C. A. Interleukin-1β maturation and release in response to ATP and nigericin. Evidence that potassium depletion mediated by these agents is a necessary and common feature of their activity. *J. Biol. Chem.* **269,** 15195–15203 (1994).
24. Perregaux, D. *et al.* IL-1β maturation: evidence that mature cytokine formation can be induced specifically by nigericin. *J. Immunol.* **149,** 1294–1303 (1992).
25. Moore, J. P., Kitchen, S. G., Pugach, P. & Zack, J. A. The CCR5 and CXCR4 coreceptors–central to understanding the transmission and pathogenesis of human immunodeficiency virus type 1 infection. *AIDS Res. Hum. Retroviruses* **20,** 111–126 (2004).
26. Schweighardt, B. *et al.* R5 human immunodeficiency virus type 1 (HIV-1) replicates more efficiently in primary CD4$^+$ T-cell cultures than X4 HIV-1. *J. Virol.* **78,** 9164–9173 (2004).
27. Grivel, J. C. & Margolis, L. B. CCR5- and CXCR4-tropic HIV-1 are equally cytopathic for their T-cell targets in human lymphoid tissue. *Nature Med.* **5,** 344–346 (1999).
28. Zhou, Y., Shen, L., Yang, H. C. & Siliciano, R. F. Preferential cytolysis of peripheral memory CD4$^+$ T cells by *in vitro* X4-tropic human immunodeficiency virus type 1 infection before the completion of reverse transcription. *J. Virol.* **82,** 9154–9163 (2008).
29. Lanzavecchia, A. & Sallusto, F. Dynamics of T lymphocyte responses: intermediates, effectors, and memory cells. *Science* **290,** 92–97 (2000).
30. Mackay, C. R. Immunological memory. *Adv. Immunol.* **53,** 217–265 (1993).
31. Sallusto, F., Lenig, D., Forster, R., Lipp, M. & Lanzavecchia, A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature* **401,** 708–712 (1999).
32. Bergsbaken, T., Fink, S. L. & Cookson, B. T. Pyroptosis: host cell death and inflammation. *Nature Rev. Microbiol.* **7,** 99–109 (2009).
33. Decker, T. & Lohmann-Matthes, M. L. A quick and simple method for the quantitation of lactate dehydrogenase release in measurements of cellular cytotoxicity and tumor necrosis factor (TNF) activity. *J. Immunol. Methods* **115,** 61–69 (1988).
34. Ventura, A. *et al.* Cre-lox-regulated conditional RNA interference from transgenes. *Proc. Natl Acad. Sci. USA* **101,** 10380–10385 (2004).
35. Baldauf, H. M. *et al.* SAMHD1 restricts HIV-1 infection in resting CD4$^+$ T cells. *Nature Med.* **18,** 1682–1687 (2012).
36. Agosto, L. M. *et al.* The CXCR4-tropic human immunodeficiency virus envelope promotes more-efficient gene delivery to resting CD4$^+$ T cells than the vesicular stomatitis virus glycoprotein G envelope. *J. Virol.* **83,** 8153–8162 (2009).
37. Boxer, M. B., Shen, M., Auld, D. S., Wells, J. A. & Thomas, C. J. A small molecule inhibitor of Caspase-1. in *Probe Reports from the NIH Molecular Libraries Program* (Bethesda Maryland, 2010).
38. Boxer, M. B. *et al.* A highly potent and selective caspase-1 inhibitor that utilizes a key 3-cyanopropanoic acid moiety. *ChemMedChem* **5,** 730–738 (2010).
39. Randle, J. C., Harding, M. W., Ku, G., Schonharting, M. & Kurrle, R. ICE/Caspase-1 inhibitors as novel anti-inflammatory drugs. *Expert Opin. Investig. Drugs* **10,** 1207–1209 (2001).
40. Stack, J. H. *et al.* IL-converting enzyme/caspase-1 inhibitor VX-765 blocks the hypersensitive response to an inflammatory stimulus in monocytes from familial cold autoinflammatory syndrome patients. *J. Immunol.* **175,** 2630–2634 (2005).
41. Maroso, M. *et al.* Interleukin-1β biosynthesis inhibition reduces acute seizures and drug resistant chronic epileptic activity in mice. *Neurotherapeutics* **8,** 304–315 (2011).
42. Vezzani, A. *et al.* ICE/caspase 1 inhibitors and IL-1β receptor antagonists as potential therapeutics in epilepsy. *Current Opinion in Investigational Drugs* **11,** 43–50 (2010).
43. Février, M., Dorgham, K. & Rebollo, A. CD4$^+$ T cell depletion in human immunodeficiency virus (HIV) infection: role of apoptosis. *Viruses* **3,** 586–612 (2011).
44. Monroe *et al.* IFI16 DNA sensor is required for death of lymphoid CD4 T cells abortively infected with HIV. *Science* http://dx.doi.org/10.1126/science.1243640 (19 December 2013).
45. Biancotto, A. *et al.* HIV-1 induced activation of CD4$^+$ T cells creates new targets for HIV-1 infection in human lymphoid tissue *ex vivo*. *Blood* **111,** 699–704 (2008).
46. Zeng, M. *et al.* Cumulative mechanisms of lymphoid tissue fibrosis and T cell depletion in HIV-1 and SIV infections. *J. Clin. Invest.* **121,** 998–1008 (2011).
47. Deeks, S. G. HIV infection, inflammation, immunosenescence, and aging. *Annu. Rev. Med.* **62,** 141–155 (2011).

Acknowledgements We thank D. N. Levy for the NLENG1 plasmid; L. A. J. O'Neill for CRID3 and parthenolide; R. Collman for the HIV-1 89.6 clone; and Vertex Pharmaceuticals for the VX-765 and VRT-043198 compounds. HIV-infected lymph node tissue was obtained from the SCOPE cohort at HIV/AIDS clinic of the San Francisco General Hospital (SFGH) Positive Health Program, with the help of R. Hoh, and M. Kerbelski. We thank W. Schecter for surgical removal of the lymph nodes from HIV-infected subjects. We thank L. Napolitano and Y. Lie from Monogram Biosciences for performing Trofile assays to determine HIV co-receptor tropism in samples of HIV-infected volunteers. The following reagents were obtained through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH: AMD3100, efavirenz and raltegravir. We thank C. Miller, director of the Gladstone Histology Core for performing the immunostaining assays and M. Cavrois, M. Gesner, and J. Tawney for assistance with flow cytometry. We also thank G. Howard and A. L. Lucido for editorial assistance; J. C. W. Carroll, G. Maki, and T. Roberts for graphics arts; and R. Givens and S. Cammack for administrative assistance. Special thanks to N. Roan for comments on the manuscript and to J. Neidleman for stimulating discussions and technical advice. We thank the NIH/NIAID for funding (R21AI102782, 1DP1036502, U19 AI0961133). Funding was also provided by the UCSF/Robert John Sabo Trust Award (G.D.) and A.P. Giannini Foundation Postdoctoral Research Fellowship (K.M.M.). We also acknowledge support from NIH P30 AI027763 (UCSF-GIVI Center for AIDS Research) for support to S.S. and Z.Y., and for Immunology Core services.

Author Contributions G.D. identified the involvement of caspase 1 and pyroptosis in lymphoid CD4 T-cell death by HIV-1, developed and designed most of the studies, collected the data and wrote the manuscript; N.L.K.G. performed IL-1β protein assays and examined VX-765 in HIV-infected tonsils; X.G. performed FLICA and shRNA analyses in HLACs; Z.Y. analysed caspase cleavage in HIV-infected cultures; K.M.M. examined caspase inhibitors and LDH release assays; O.Z. tested caspase inhibitors, type-I IFN, and pro-IL-1β expression; P.W.H. and H.H. provided HIV-infected lymphoid node from surgeries of SCOPE cohort patients at HIV/AIDS clinic of the San Francisco General Hospital (SFGH); I.M.-A. provided reagents and tissues; S.S. coordinated lymph node biopsies; W.C.G. supervised all of these studies and participated in the preparation of the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.C.G. (wgreene@gladstone.ucsf.edu).

METHODS

Preparation of HIV-1 virions. To generate replication-competent viruses, pNL4-3, pNLENG1 or 89.6, proviral expression DNAs were transfected into 293T cells by the calcium phosphate method. The medium was replaced after 16 h. After 48 h, the supernatants were collected and clarified by sedimentation, the virions were concentrated by ultracentrifugation, and were stored at −80 °C in 100% fetal bovine serum. All viral stocks were quantitated by measuring p24 Gag levels by ELISA (1 ng of $p24^{gag}$ equals approximately 2×10^6 viral particles). The R5-tropic GFP reporter virus (pBRNL43_005pf135(R5)nef+_IRES_eGFP) was derived from the pNLENG1 clone replaced with gp120 V3 loop sequence of R5-tropic HIV primary isolates as previously described[48].

Culture and infection of HLACs. Human tonsil or splenic tissues were obtained from the National Disease Research Interchange and the Cooperative Human Tissue Network and processed as previously described[11]. HLACs were infected with HIV-1 in 96-well V-bottomed polystyrene plates by spinoculation of 80 ng p24 Gag of HIV particles with 1×10^6 cells in a total of 100 μl per well. Cells were chilled on ice for 15 min and HIV-1 was then added to each well and mixed with cold cells. Virions and cells were subjected to high-speed centrifugation (1200*g*) for 2 h at 4 °C. This step promotes high-level attachment of virions to target cell membranes. Immediately after centrifugation, cells were cultured at 37 °C as a pellet to facilitate synchronized fusion of the attached viruses. After 10 h of incubation and establishment of productive infection, the indicated drugs were added. Because splenic cells are extremely refractory to HIV infection, we modified the infection system by overlaying splenic HLAC cells on a monolayer of 293T cells that had been transfected with HIV-1 proviral clones. Analysis of CCR5-expressing CD4 T-cell death was similarly performed using 293T transfected with the R5-tropic 81A strain of HIV-1. We also used this method for assays using shRNA-infected HLACs. The 293T cells were transfected with 50 μg of HIV-1 DNA in a 24-well plate. After 12 h, 293T cells were overlayed with 4×10^6 HLACs per well in RPMI media in the presence of the indicated drugs. Virus-producing 293T cells directly interact with overlaid target HLACs. After 24–72 h, the HLAC suspensions were collected from wells and analysed by flow cytometry. Unless otherwise stated, drugs were used at the following concentrations: AMD3100 (250 nM); efavirenz (100 nM); nigericin (8–10 μM); staurosporine (50 nM); Ac-YVAD-CMK, Z-WEHD-FMK, Z-DEVD-FMK, Z-VAD FMK, Z-VEID-FMK, Z-VAD FMK, Z-IETD-FMK or Z-FA-FMK (all 50 μM) (100 μM was determined to be the maximal concentration of these caspase inhibitors that is not associated with toxicity); VX-765 (10 μM); VRT-043198 (10 μM); necrostatin (5 μM); CRID3 (50 μM); parthenolide (10 μM); glyburide (20 μM); glimepiride (20 μM) (20 μM of glyburide and glimepiride was determined to be the maximal drug concentration that does not induce toxicity). VX-765 was added to the cultures 4 hours before infection to allow absorption and processing by the cells.

FACS analysis and gating strategy. HLACs were washed in FACS buffer (PBS supplemented with 2 mM EDTA and 2% fetal bovine serum), stained with PE-conjugated anti-CD4, PerCP-conjugated anti-CD19, and APC-conjugated anti-CD8 (all from BD Pharmingen) and fixed in 2% paraformaldehyde. For analysis of CCR5-expressing CD4 T cells, HLACs were stained with 1:3 dilutions of mouse anti-human CCR5 (BD Pharmingen, clone 2D7/CCR5) on ice for 3 h. In isolated CD4 T-cell cultures a standard number of fluorescent beads (Flow-Count Fluorospheres, Beckman Coulter) were added to each cell-suspension sample before data acquisition. Data were collected on a FACSCalibur (BD Biosciences) and analysed with FlowJo software (Treestar). The percentages of viable CD4 T cells were defined by sequential gating beginning with forward scatter versus side scatter to select live lymphocytes, then calculating the number of CD4 T cells divided by the number of CD8 T cells, or by normalization based on the number of fluorescent beads acquired by volume.

Protein analysis, LDH assay, IFN inhibition and intracellular caspase stainings. In order to stimulate the processing and secretion of IL-1β, CD4 T cells were isolated from HLACs by positive selection and treated with 8–10 μM nigericin (Sigma) for 12 h at 37 °C. The potassium ionophore nigericin mediates an electroneutral exchange of intracellular K^+ ions for extracellular protons, providing a second inflammatory stimulus, which results in the NLRP3-mediated activation of caspase 1 (ref. 19). In order to assess the processing and secretion of IL-1β in infected CD4 T cells, CD4 T cells were isolated from HLACs as described above, spinoculated with or without NL4-3 (80 ng p24 Gag per 1×10^6 cells) and the indicated drugs as described in the figures. For cytoplasmic pro-IL-1β (Fig. 2b) and other intracellular protein analyses, cells were washed in PBS and immediately lysed in cell extraction buffer (Life Technologies) containing a protease inhibitor cocktail (Roche). For NLRP3 detection cells were lysed using digitonin lysis buffer (digitonin 0.5%, 20 mM Tris-HCl (pH 7.4), 150 mM NaCl) with the addition of a protease inhibitor cocktail (Roche). Lysates were subjected to SDS–PAGE protein analysis using mouse anti human IL-1β antibody (R&D systems clone 8516, catalogue number MAB201), which recognizes the pro- as well as the cleaved form of IL-1β (Fig. 2a, b). For analysis of secreted IL-1β (Figs 2a, c, and 3g), cells were cultured in RPMI 1640 supplemented with 5% heat-inactivated fetal bovine serum. Supernatants were collected 3–5 days after infection with HIV-1 or 12 h after treatment with nigericin, filtered through 0.22-μm filter plates (Millipore) and subjected to SDS–PAGE protein analysis using rabbit polyclonal anti-human IL-1β (Abcam, catalogue number ab2105), which primarily recognizes the cleaved form of IL-1β, or assessed for release of cytoplasmic lactate dehydrogenase (LDH) as previously described[33]. For SDS–PAGE immunoblotting analysis, Bio-Rad Criterion 15% pre-cast Tris-HCL gels were used. Gels were wet transferred onto PVDF membranes (Bio-Rad) at maximum current for 3 h at 4 °C and then blocked in 5% non-fat milk for 1 h at room temperature. Primary antibodies were incubated overnight at 4 °C and secondary antibodies for 1 h at room temperature. Additional primary antibodies used for SDS–PAGE analysis were 1:1,000 rabbit anti-caspase 1 p10 (clone c-20, Santa Cruz, catalogue number SC-515), 1:1,000 rabbit anti-caspase 3 (clone 8G10, Cell Signaling, catalogue number 9665S), 1:1,000 mouse anti NLRP3 (Abcam, catalogue number ab17267), 1:1,000 rabbit polyclonal anti-human ASC (Imgenex, catalogue number IMG-5662), 1:100 Phospho-Stat1 (ser727, Cell Signaling, catalogue number 9177), and 1:10,000 of the mouse monoclonal anti-β-actin (Sigma, catalogue number A5316). The secondary antibody used was 1:5,000 anti-rabbit secondary (Thermo Scientific, catalogue number 32460) or 1:5,000 anti-mouse secondary (Thermo Scientific, catalogue number 32430) developed using a 1:4 dilution of SuperSignal West Femto substrate (Thermo Scientific). To neutralize interferon receptors in HLACs, cultures were added with 1–5 μg of anti-interferon-α/β receptor chain 2 antibody, clone MMHAR-2 (Millipore). To determine intracellular activation of specific caspases, fluorescent labelled inhibitors of caspases (FLICA) probe assays (ImmunoChemistry Technologies) were performed. Each FLICA probe contains a 3 or 4 amino acid sequence targeted by a specific activated caspase. There is no interference from pro-caspases or the inactive form of the enzymes[17]. FLICA probes were added directly to the cell culture media, incubated for 15 min at 37 °C, and washed five times with PBS supplemented with 2 mM EDTA and 2% fetal bovine serum. FLICA probes are cell-permeable and covalently bind to the active forms of specific caspases. After washing, FLICA fluorescent signal is specifically retained within cells containing the appropriate active form of the caspase while the reagent is washed away in cells lacking the appropriate active caspase.

Production and infection of Vpx-VLPs and shRNA-coding HIV LV particles. SIVmac 251 virus-like particles for Vpx delivery (Vpx-VLPs) were produced using the pSIV3$^+$ plasmid, provided by A. Cimarelli[49]. These Vpx-VLP particles are non-infectious as they do not contain any viral genetic material, but they are used to transiently deliver Vpx into target cells where it promotes degradation of SAMHD1 thereby rendering the cells permissive to HIV LV infection[50]. Rather than using the typical pseudotyping method with VSV-G glycoprotein, we pseudotyped the Vpx-VLPs with the CXCR4-tropic Env of HIV-1, which supports efficient fusion of viral particles to quiescent CD4 T lymphocytes[36]. For production of Vpx-VLPs 293 T-cells were co-transfected with 8 μg pSIV3$^+$ and 2 μg CXCR4-tropic Env (gp160)-encoding plasmid. The amount of lentiviral particles was determined by SIV p27 Gag ELISA assay. shRNA-coding vectors were cloned using a modified version of the pSicoR (plasmid for stable RNA interference, conditional) lentiviral vector[34], which encodes an mCherry reporter driven by an EF-1α promoter (pSicoR-MS1)[51]. To generate shRNA lentiviral particles, 293T cells were co-transfected with 10 μg pSicoR-mCherry shRNA constructs, 9 μg HIV-based packaging construct NL4-3 8.91 (ref. 52), and 2 μg CXCR4-tropic Env (gp160)-encoding plasmids. Cells were transfected using the standard phosphate calcium transfection protocol[53]. The lentiviral particle stocks were quantitated by HIV p24 Gag ELISA assay (1 ng of p24 Gag equals approximately 2×10^6 viral particles).

To achieve productive infection of shRNA-encoding LV particles, complete HLACs or isolated lymphoid CD4 T cells were initially challenged with Vpx-VLPs, followed by a second infection with an shRNA-coding LV of interest after 24 h. This sequential infection strategy allowed Vpx to establish an optimal permissive state within the target cells at the time when the shRNA LV infection was performed. To facilitate a synchronized delivery of Vpx and fusion of shRNA LV particles, cells and particles were subjected to high-speed spinoculation at each step. To assess the efficiency of gene silencing by the shRNA-coding vectors (Fig. 3d), highly infectious SupT1 were infected with shRNA LV (without prior Vpx-VLP infection), and were subjected to protein analysis after 48 h.

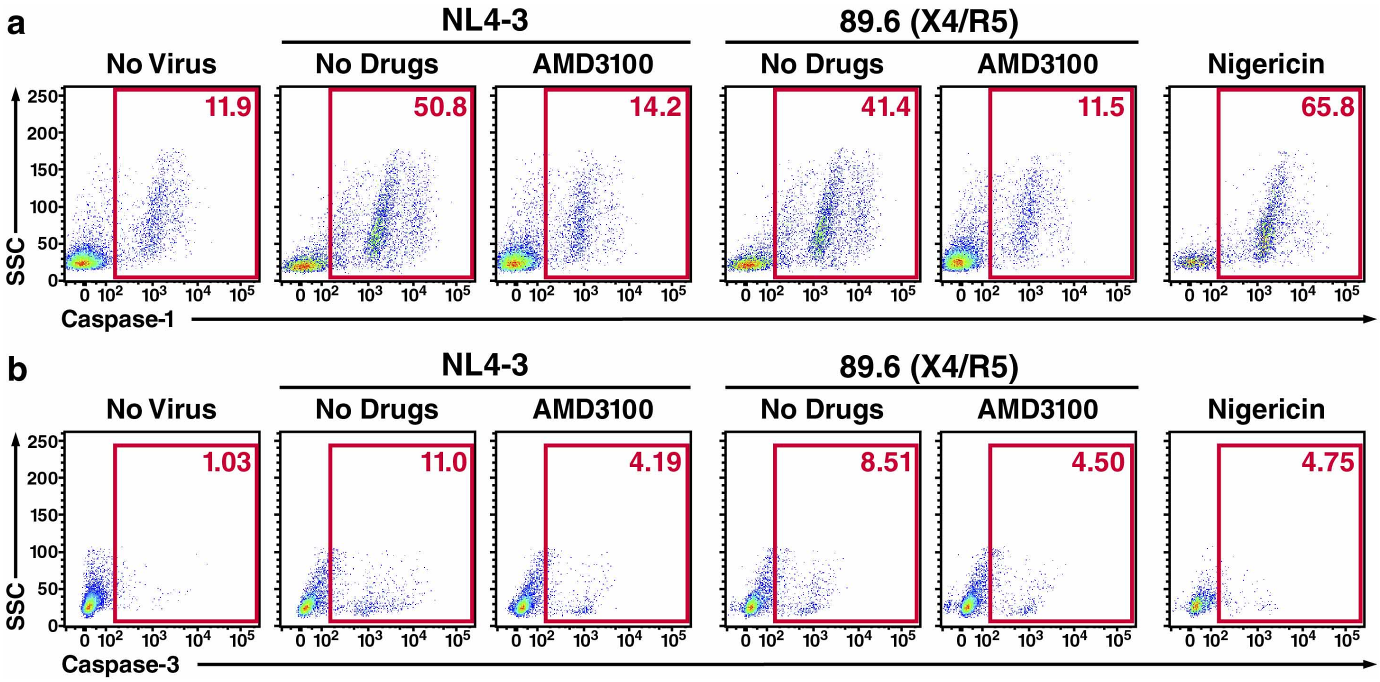
For cloning of caspase 1 coding shRNA vector the following oligonucelotides were used: sense: 5′-TACACGTCTTGCTCTCATTATTCAAGAGATAATGAGAGCAAGACGTGTTTTTTTC-3′; antisense: 5′-TCGAGAAAAAAACACGTCTTGCTCTCATTATCTCTTGAATAATGAGAGCAAGACGTGTA-3′. For cloning of caspase 3 coding shRNA vector the following oligonucleotides were used: sense: 5′-TAAAGGTGGCAACAGAATTTTTCAAGAGAAAATTCTGTTGCCACCTTTTTTTTC-3′; antisense: 5′-TCGAGAAAAAAAAGGTGGCAACAGAATTTTCTCTTGAAAAATTCTGTTGCCACCTTTA-3′. For cloning of ASC-coding shRNA vector the following oligonucelotides were used: sense: 5′-TGAA

GCTCTTCAGTTTCACATTCAAGAGATGTGAAACTGAAGAGCTTCTTTTTTC-3′; antisense: 5′-TCGAGAAAAAAGAAGCTCTTCAGTTTCACATCTCTTGAATGTGAAACTGAAGAGCTTCA-3′. For cloning of NLRP3-coding shRNA vector the following oligonucelotides were used: sense: 5′-TGAAATGGATTGAAGTGAAATTCAAGAGATTTCACTTCAATCCATTTCTTTTTTC-3′; antisense: 5′-TCGAGAAAAAAGAAATGGATTGAAGTGAAATCTCTTGAATTTCACTTCAATCCATTTCA-3′.

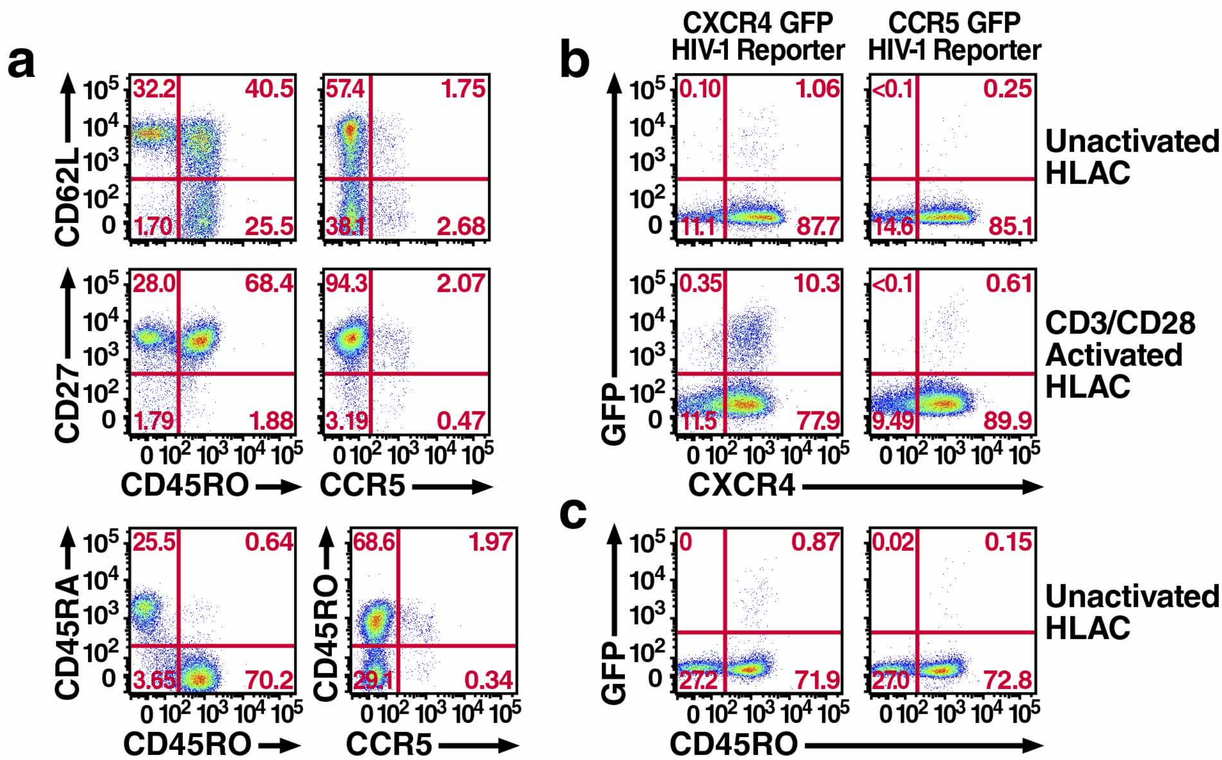
Tissue samples. HIV-infected lymph node tissues were obtained from patients participating in the SCOPE cohort at HIV/AIDS clinic of the San Francisco General Hospital (SFGH) Positive Health Program. All tissues were obtained with full consent from the patients and under a protocol fully approved by the Committee on Human Research at UCSF. For the results presented, an inguinal lymph node was harvested from two different HIV-infected patients: A 50-year-old immunosuppressed, untreated R5-tropic HIV-1-infected subject in the chronic phase of disease. This individual exhibited a viral load of 87,756 RNA copies per ml, and CD4 T-cell count of 227 cells per μl. A 41-year-old African-American male, infected with an R5-tropic strain of HIV-1, had been on intermittent anti-retroviral therapy between 2004–2009 and stopped anti-retroviral therapy in 2009. This individual exhibited a viral load of 30,173 RNA copies per ml, and a CD4 T-cell count of 259 cells per μl. The fresh specimens were immediately fixed with 4% PFA and subjected to immunostaining analysis. Sections of the HIV-infected lymph node and of a fresh human tonsil were processed in parallel and analysed for the indicated markers. IRB approval number 10-03606 with study title: the use of lymph node biopsies to support HIV pathogenesis studies.

Tissue preparation and immunohistochemistry. Five-micron sections were cut from formalin-fixed paraffin-embedded tissue blocks and mounted on X-tra microscope slides (Leica Microsystems). Specimens were stepwise deparaffinized in xylene and rehydrated in descending alcohols to water. Endogenous peroxidase activity was blocked by incubation in 0.3% hydrogen peroxide (Sigma Chemicals) in PBS for 15 min. Antigen retrieval was performed by microwaving the sections in 10 mM citrate buffer, pH 6.0. Sections were then blocked in the secondary antibody host's normal serum (Vector Labs; horse S-2000, goat S-1000, catalogue number S5000). The following primary antibodies were diluted in PBS with 0.1% bovine serum albumin (BSA) and applied to the slides overnight at 4°C: monoclonal mouse anti-human CD3 (1:100, Clone F7.2.38 Dako, catalogue number M725429-2), monoclonal rabbit anti-human CD11c (1:100, clone EP1347Y, Abcam catalogue number ab52632), monoclonal mouse anti-human Ki-67 (1:100, clone MIB1, Dako catalogue number M724029-2), monoclonal mouse anti-HIV p24 Gag (1:50, clone KaI-1, Dako Cytomation catalogue number M0857), rabbit anti-human cleaved caspase 3 (1:300, Cell Signalling Technology catalogue number 9661), goat anti-p20 subunit of active human caspase 1 (1:200, clone c15, Santa Cruz Biotechnology catalogue number sc-1780), rabbit anti-human against bioactive 17-kDa IL-1β (1:100, Abcam catalogue number ab2105), and annexin V (1:50, Abcam catalogue number EPR3979). The following day sections were washed in 0.05% Tween-20 in PBS followed by incubation with Vector laboratories biotinylated secondary IgG antibodies diluted 1:200 in PBS for 30 min at room temperature (donkey anti-mouse BA-2000, goat anti-rabbit BA-1000, rabbit anti-goat BA-5000). Slides were then rinsed in 0.05% Tween-PBS, and incubated in streptavidin horseradish peroxidase complex at a 1:200 dilution in PBS for 30 min at room temperature (Vector Laboratories catalogue number SA-5004). Specimens were rinsed in 0.05% Tween-20 in PBS then incubated with 3,3-diaminobenzadine (DAB) chromogenic substrate (Sigma Chemicals) using hydrogen peroxide as a substrate (Sigma Chemicals) for 10 min. Sections were counterstained in haematoxylin dehydrated through graded alcohols, cleared in xylene and mounted in depex.

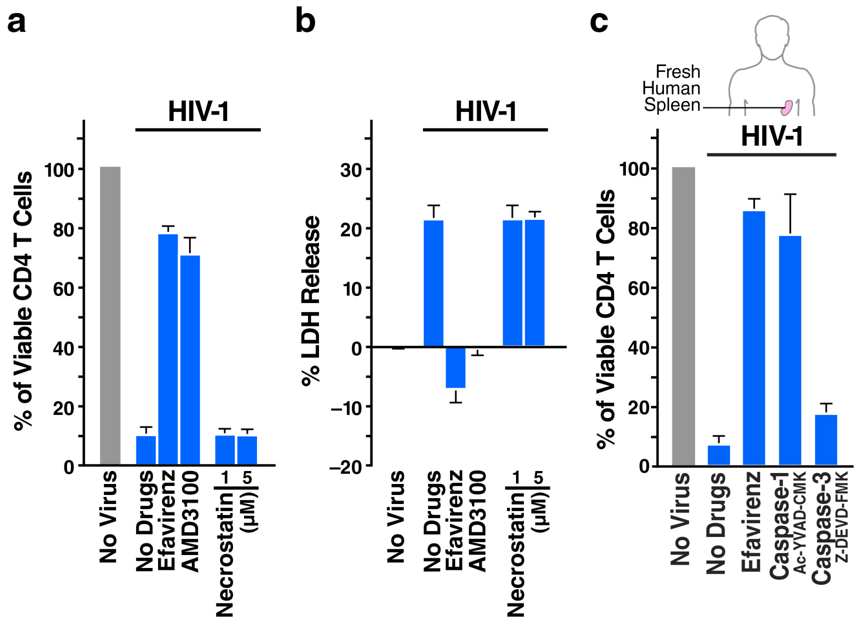
48. Papkalla, A., Munch, J., Otto, C. & Kirchhoff, F. Nef enhances human immunodeficiency virus type 1 infectivity and replication independently of viral coreceptor tropism. *J. Virol.* **76,** 8455–8459 (2002).
49. Goujon, C. *et al.* With a little help from a friend: increasing HIV transduction of monocyte-derived dendritic cells with virion-like particles of SIV(MAC). *Gene Ther.* **13,** 991–994 (2006).
50. Laguette, N. *et al.* SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature* **474,** 654–657 (2011).
51. Wissing, S., Montano, M., Garcia-Perez, J. L., Moran, J. V. & Greene, W. C. Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells. *J. Biol. Chem.* **286,** 36427–36437 (2011).
52. Grivel, J. C. *et al.* HIV-1 pathogenesis differs in rectosigmoid and tonsillar tissues infected ex vivo with CCR5- and CXCR4-tropic HIV-1. *AIDS* **21,** 1263–1272 (2007).
53. Wigler, M., Pellicer, A., Silverstein, S. & Axel, R. Biochemical transfer of single-copy eucaryotic genes using total cellular DNA as donor. *Cell.* **14,** 725–731 (1978).
54. Doranz, B. J. *et al.* A dual-tropic primary HIV-1 isolate that uses fusin and the beta-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. *Cell* **85,** 1149–1158 (1996).
55. De Rosa, S. C., Herzenberg, L. A., Herzenberg, L. A. & Roederer, M. 11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity. *Nature Med.* **7,** 245–248 (2001).
56. Brenchley, J. M. *et al.* CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *J. Exp. Med.* **200,** 749–759 (2004).
57. Bleul, C. C., Wu, L., Hoxie, J. A., Springer, T. A. & Mackay, C. R. The HIV coreceptors CXCR4 and CCR5 are differentially expressed and regulated on human T lymphocytes. *Proc. Natl Acad. Sci. USA* **94,** 1925–1930 (1997).
58. Gondois-Rey, F. *et al.* Segregation of R5 and X4 HIV-1 variants to memory T cell subsets differentially expressing CD62L in *ex vivo* infected human lymphoid tissue. *AIDS* **16,** 1245–1249 (2002).
59. Penn, M. L., Grivel, J. C., Schramm, B., Goldsmith, M. A. & Margolis, L. CXCR4 utilization is sufficient to trigger CD4+ T cell depletion in HIV-1-infected human lymphoid tissue. *Proc. Natl Acad. Sci. USA* **96,** 663–668 (1999).
60. Cho, Y. S. *et al.* Phosphorylation-driven assembly of the RIP1-RIP3 complex regulates programmed necrosis and virus-induced inflammation. *Cell* **137,** 1112–1123 (2009).
61. Pitha, P. M. Innate antiviral response: role in HIV-1 infection. *Viruses* **3,** 1179–1203 (2011).
62. Samuel, C. E. Antiviral actions of interferons. *Clinical Microbiology Reviews* **14,** 778–809 (2001).
63. Lamkanfi, M. & Dixit, V. M. The inflammasomes. *PLoS Pathog.* **5,** e1000510 (2009).
64. Coll, R. C. & O'Neill, L. A. The cytokine release inhibitory drug CRID3 targets ASC oligomerisation in the NLRP3 and AIM2 inflammasomes. *PLoS ONE* **6,** e29539 (2011).
65. Juliana, C. *et al.* Anti-inflammatory compounds parthenolide and Bay 11-7082 are direct inhibitors of the inflammasome. *J. Biol. Chem.* **285,** 9792–9802 (2010).
66. Lamkanfi, M. *et al.* Glyburide inhibits the Cryopyrin/Nalp3 inflammasome. *J. Cell Biol.* **187,** 61–70 (2009).
67. MacDonald, K. P. *et al.* Characterization of human blood dendritic cell subsets. *Blood* **100,** 4512–4520 (2002).
68. Merad, M., Sathe, P., Helft, J., Miller, J. & Mortha, A. The dendritic cell lineage: ontogeny and function of dendritic cells and their subsets in the steady state and the inflamed setting. *Annu. Rev. Immunol.* **31,** 563–604 (2013).
69. Descours, B. *et al.* SAMHD1 restricts HIV-1 reverse transcription in quiescent CD4+ T-cells. *Retrovirology* **9,** 87 (2012).



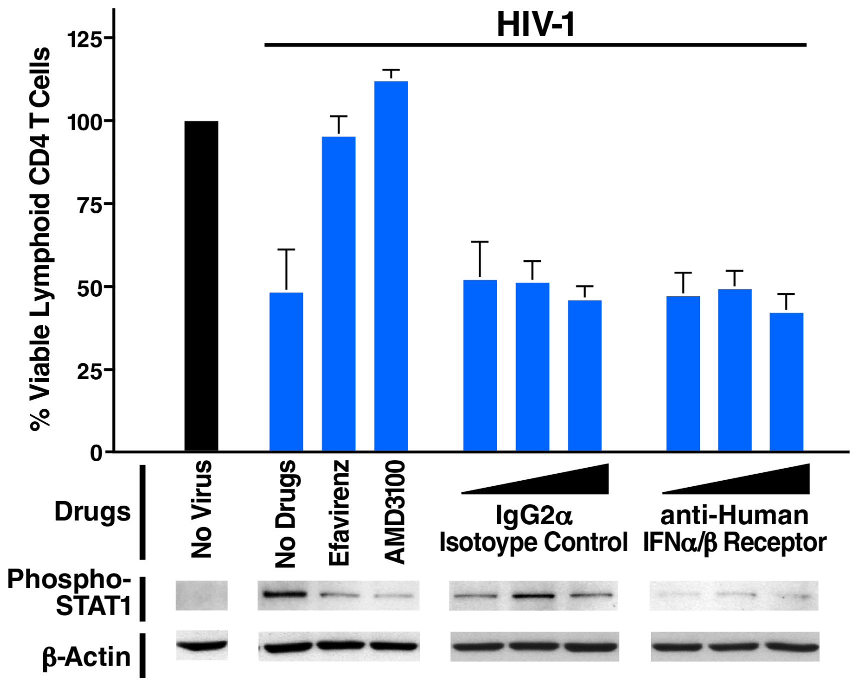
Extended Data Figure 1 | Extensive caspase 1 activation in dying lymphoid CD4 T cells infected with either NL4-3 or a primary HIV-1 isolate. **a**, Dying CD4 T cells activate caspase 1. HLACs were infected with NL4-3 or with a primary HIV-1 isolate 89.6 obtained from a mixed PBMC culture from an AIDS patient. The 89.6 viral isolate replicates to high titres in primary human cells such as macrophages and lymphocytes. It is highly cytopathic and utilizes both CCR5 and CXCR4 as co-receptors (dual-tropic)[18,54]. Infected cells were treated either with no drugs or with AMD3100 (250 nM) entry inhibitor, as indicated. Caspase 1 activity was determined by flow cytometry using FLICA 12 h after treatment with nigericin (10 μM) or 3 days after infection with HIV. Notably, equivalent levels of caspase 1 activation were observed in CD4 T cells infected with NL4-3 or 89.6 HIV-1 isolate. AMD3100 prevented caspase 1 activity with both viruses, indicating the abundant presence of CXCR4-expressing target CD4 T cells in these cultures. **b**, Low levels of caspase 3 activity in dying CD4 T cells. The same cultures as in (**a**) were tested for caspase 3 activity using FLICA. Compared to caspase 1, infections with NL4-3 and 89.6 HIV-1 isolate induced low levels of caspase 3 activation in dying CD4 T cells. No caspase 3 activation was observed in cells treated with nigericin, which signals the NLRP3 inflammasome to activate caspase 1 (ref. 19), indicating a specific recognition of caspase 1 and caspase 3 activity by the FLICA probes. These data are the representative results of four independent experiments performed in tonsil cells isolated from four different donors.



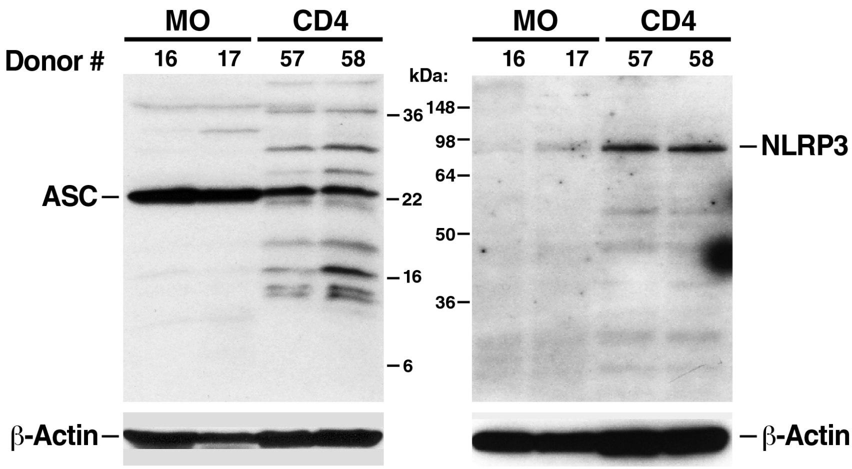
Extended Data Figure 2 | Resting CD4 T cells from tonsil include both naive and memory subsets. **a**, CD4 T lymphocytes in lymphoid tissues contain a large population of central memory cells. To identify the sub-populations of CD4 T cells in human tonsil histocultures, we examined the expression pattern of CCR5, CD45RA, CD45RO, CD62L and CD27. Central memory CD4 T cells (T_{CM}) are characterized by expression of $CD45RO^+$/$CD62L^+$ or $CD45RO^+$/$CD27^+$ [29,31,55,56]. T_{CM} lack effector function and constantly travel through the lymph nodes in large quantities for antigen sampling, whereas effector memory cell (T_{EM}) mainly migrate to peripheral tissues[29–31]. Analysis of these surface markers revealed at least three distinct maturation phenotypes. The majority of CD4 T lymphocytes exhibit a memory phenotype as determined by surface expression of CD45RO, among them more than two-thirds were found to be central memory cells ($CD45RO^+$/$CD62L^+$ and $CD45RO^+$/$CD27^+$). Similarly, a large population of CCR5-expressing CD4 T cells was found to have central memory phenotype ($CCR5^+$/$CD62L^+$ and $CCR5^+$/$CD27^+$). These findings are in accordance with previous studies in primary human lymphoid cultures[12,57,58]. **b**, **c**, Memory lymphoid CD4 T cells represent preferential targets for productive infection by both the R5- and X4-tropic strains of HIV-1. To determine whether cell maturation influences susceptibility for productive infection, we measured the levels of productive infection using GFP reporter viruses harbouring either an X4-tropic or R5-tropic Env of HIV-1. Except for their select V3 loop envelope determinants, both reporters were derived from the same bicistronic Nef-IRES-GFP clone which produces fully replication-competent viruses[16]. Interestingly, productive infection of both X4-tropic or R5-tropic viral strains was detected in CXCR4-expressing cells, indicating that the CXCR4 co-receptor is equally present on CCR5-expressing cells, as was previously shown[12,57–59]. Memory CD4 T cells ($CD45RO^+$) were selectively productively infected in cultures infected with either X4-tropic or R5-tropic reporter virus. Similar findings were found in infected cultures activated with CD3/CD38 beads to achieve higher rates of infection. Among the memory CD4 T cells, T_{EM} cells became productively infected in higher quantities than T_{CM} (not shown). These data are the representative results of six independent analyses performed in tonsil cells isolated from six different donors.



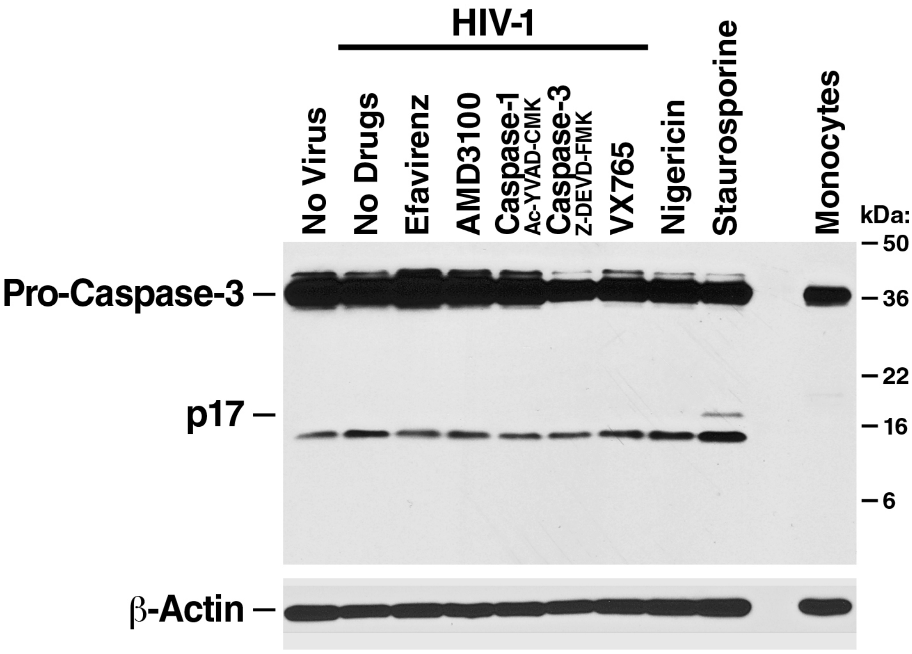
Extended Data Figure 3 | Necrostatin-1 does not prevent lymphoid CD4 T-cell death and cell lysis in HIV-infected cultures. **a**, **b**, Necrostatin was tested at 1 or 5 μM, a concentration that yields maximal inhibition without inducing toxicity (not shown). Pyroptosis shares cell death features with necroptosis which similarly leads to the release of intracellular contents into the extracellular space[9]. To test whether cell death involves necrotic signalling, we treated HIV-infected CD4 T cells with necrostatin, a specific inhibitor of RIP1, whose kinase activity is essential for programmed necroptosis to occur[60]. Concentrations of necrostatin that block necroptotic signalling (not shown) did not inhibit CD4 T-cell depletion in HIV-infected cultures (**a**), and did not prevent the release of intracellular contents into the culture medium, as indicated by LDH activity in the supernatants (**b**). Thus, although pyroptosis shares features with necroptosis, these data demonstrate that the signalling pathways linking caspase 1 activation to CD4 T-cell death are specific. Together, these findings indicate that the CD4 T-cell depletion and release of cytoplasmic contents in HIV-infected lymphoid cultures reflects pyroptosis rather than apoptosis or necroptosis. Error bars represent s.e.m. of at least three independent experiments using tonsil cells from at least three different donors. **c**, Caspase 1 inhibitors prevent CD4 T-cell death in HIV-infected splenic tissues. Splenic HLACs were cultured with no virus or were infected with HIV-1. The HIV-infected cultures were treated as indicated, either with no drugs, efavirenz (100 nM), the caspase 1 inhibitor Ac-YVAD-CMK (50 μM), or the caspases-3 inhibitor Z-DEVD-FMK (50 μM). After 4 days, viable CD4 T cells were counted by flow cytometry. Viable CD4 T cells are presented as the percentage remaining live CD4 T cells using CD8 T cells to normalize each HIV-infected or uninfected culture. Error bars represent s.e.m. from four independent experiments using tonsil cells isolated from four different donors.



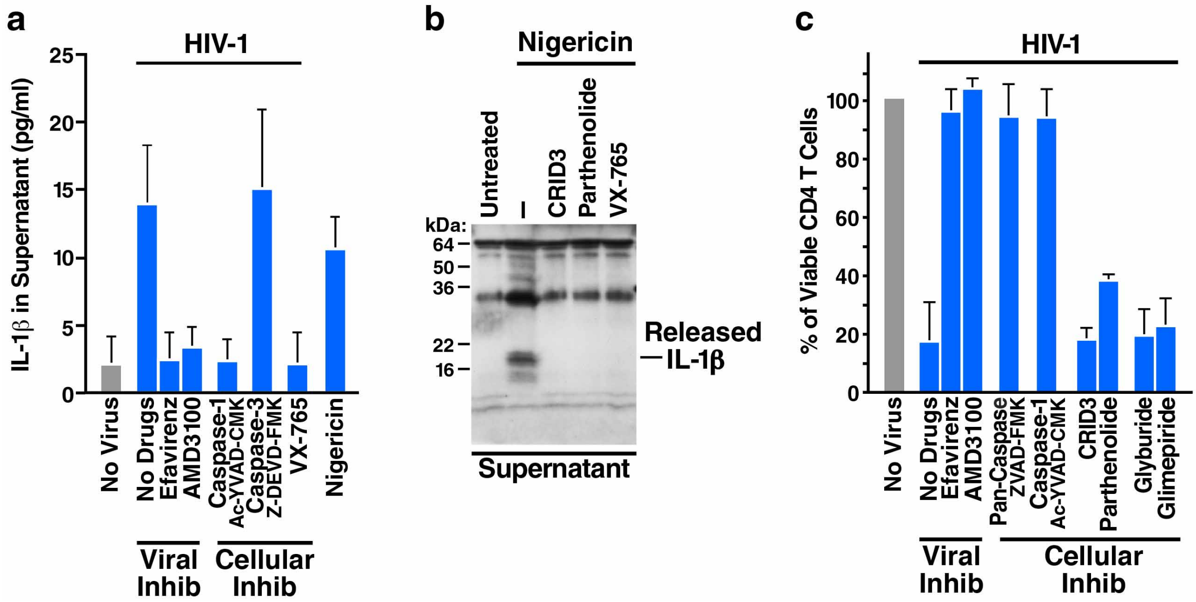
Extended Data Figure 4 | Induction of type-I interferon is not required to trigger a death response in HIV-infected lymphoid CD4 T cells. HIV-1 infections induce type-I interferon *in vitro* and *in vivo*[61]. To test the involvement of this antiviral response in modulating CD4 T-cell death, isolated CD4 T cells were infected with HIV-1 in the presence of neutralizing antibodies against the human interferon alpha receptor (IFNAR2), which blocks biological action of type I interferons. To determine the state of interferon signalling, cells were analysed in parallel for the presence of tyrosine-phosphorylated STAT1, which plays a central role in mediating type-I IFN-dependent biological responses, including induction of an antiviral state[62]. Phosphorylated STAT1 readily appeared in HIV-infected CD4 T cells, but not in HIV-infected cells treated with efavirenz (100 nM), AMD3100 (250 nM) or anti-IFNAR2 neutralizing antibodies (1–5 μg ml^{-1}). Notably, blocking interferon signalling with anti-IFNAR2 neutralizing antibodies did not prevent the death of CD4 T cells by HIV-1, although tyrosine phosphorylation of STAT1 was inhibited indicating effectiveness of the antibody blockade. The data suggest that this antiviral IFN induction is not critical to the onset of the innate immune death response leading to caspase 1 activation and pyroptosis. Error bars represent s.e.m. from three independent experiments using tonsil cells from three different donors.



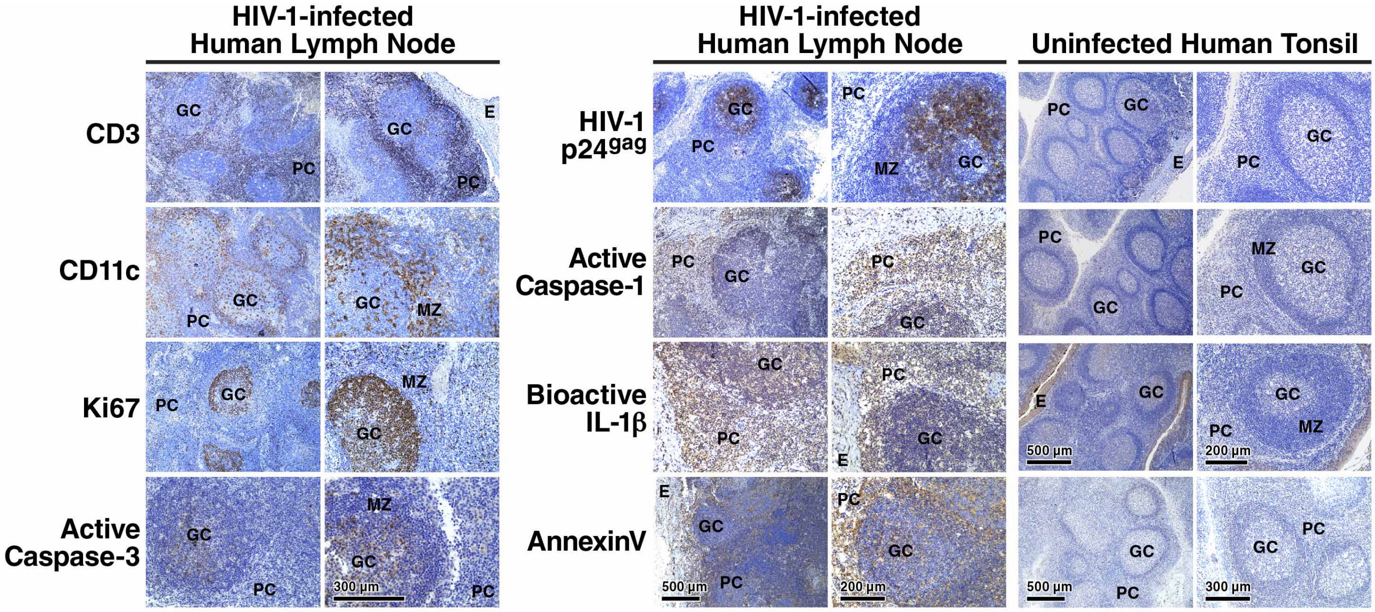
Extended Data Figure 5 | Lymphoid CD4 T cells express detectable levels of ASC and NLRP3 relative to blood-derived monocytes. The bipartite adaptor protein ASC (PYCARD) plays a central role in the interaction between (NOD)-like receptor and caspase 1 in inflammasome complexes[63]. Lymphoid CD4 T cells are primed to mount such inflammatory responses, and constitutively express high levels of cytoplasmic pro-IL-1β, but also ASC and NLRP3, compared to blood-derived monocytes. CD4 T lymphocytes express constitutive levels of NLRP3. In contrast to lymphocytes, monocytes require stimulation with TLR ligands such as LPS to induce NLRP3 expression[21]. Thus, the release of intracellular 5′-ATP by pyroptotic CD4 T cells may provide a second inflammatory stimulus to induce activation of caspase 1 by the NLRP3 inflammasome in nearby CD4 T cells that are already primed as reflected by their high levels of ASC, NLRP3 and pro-IL-1β expression. Thus, pyroptosis activated initially by HIV may result in cascade of new rounds of pyroptosis in primed CD4 T cells by the repeated release of intracellular ATP in a virus-independent manner. Such an 'auto-inflammation' scenario could result in persistent rounds of pyroptosis, chronic inflammation and loss of CD4 T cells even when viral loads are reduced by antiretroviral therapy (ART).



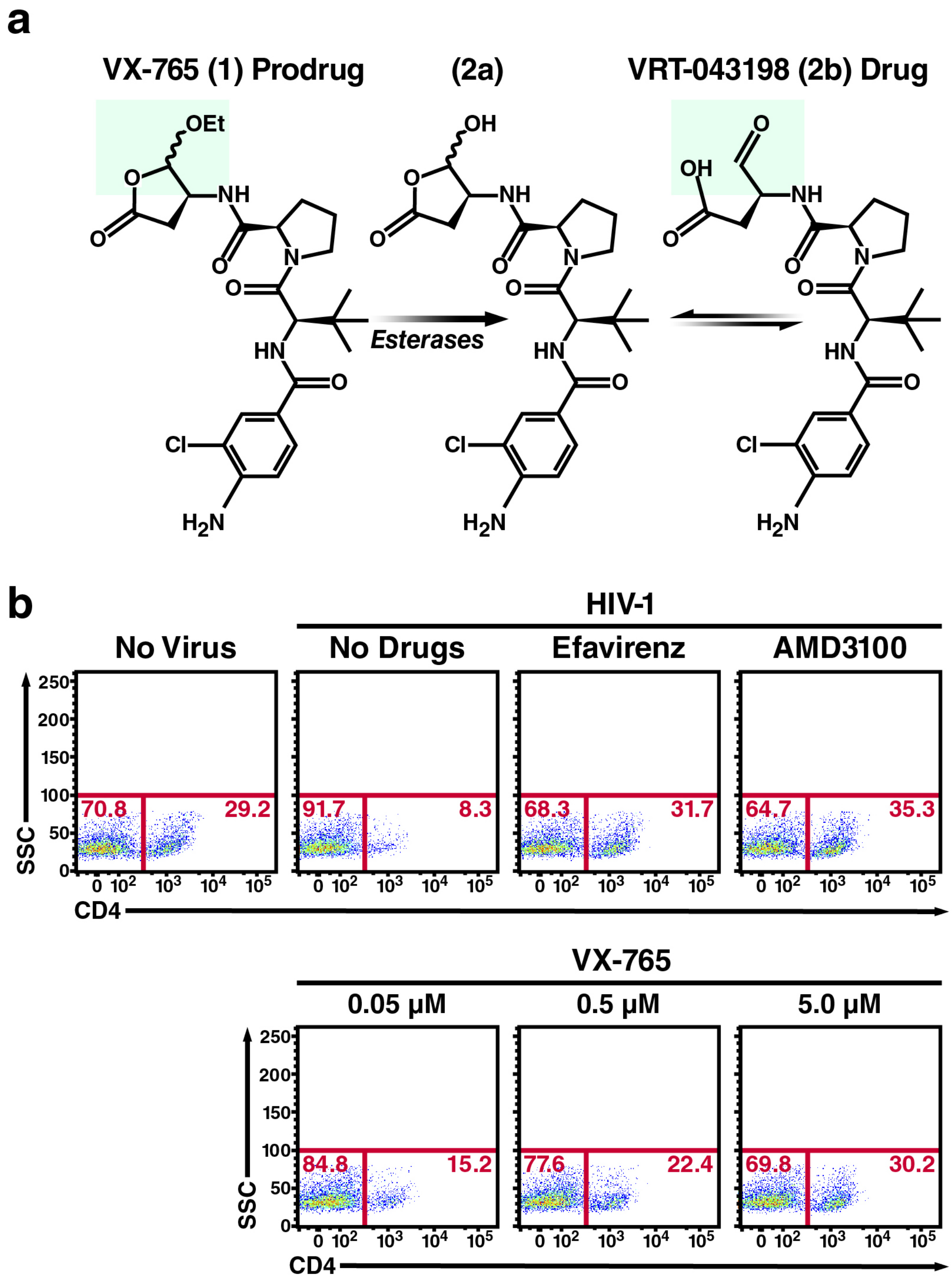
Extended Data Figure 6 | Low levels of caspase 3 activation in HIV-infected lymphoid CD4 T-cell cultures. Although the endogenous levels of pro-caspase 3 and pro-caspase 1 expression are similar in lymphoid CD4 T cells, caspase 3 activation in these cells was markedly less abundant after infection with HIV-1, compared to caspase 1. These data are in accord with our findings using fluorescently labelled inhibitor of caspases (FLICA) probes in cultures infected with a GFP reporter HIV-1. In these cultures, the majority of CD4 T cells were abortively infected and showed activation of intracellular caspase 1. No caspase 1 activity was observed in productively infected cells (Fig. 1b). In sharp contrast, caspase 3 activity in these cultures was markedly less abundant, and specifically occurred in productively infected, but not in non-productively infected cells (Fig. 1c). These data are the representative results of three independent experiments performed in tonsillar CD4 T cells isolated from three different donors.



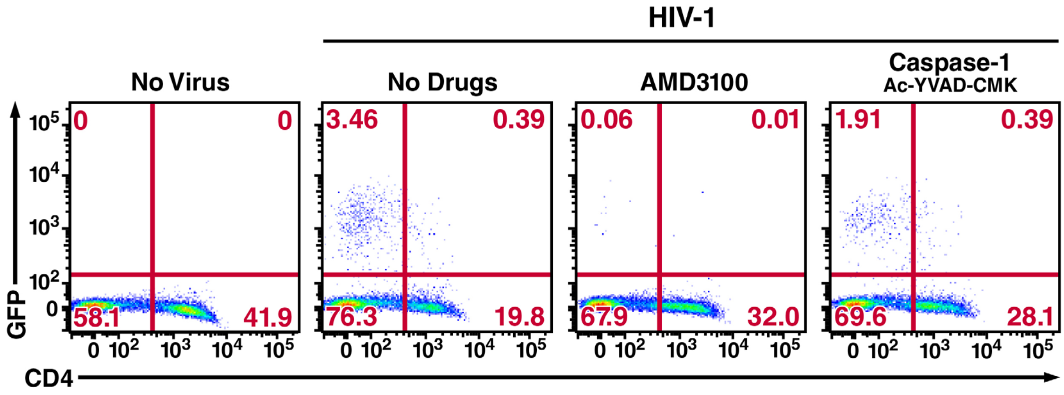
Extended Data Figure 7 | Inhibitors of caspase 1, but not of NLRP3, prevent CD4 T-cell death by HIV-1. **a**, Quantitative evaluation of bioactive IL-1β secreted in HIV-infected CD4 T-cell cultures using ELISA. Isolated tonsillar CD4 T cells were left uninfected or infected with HIV in the presence of the indicated drugs. Four days after infection, supernatants were filtered through 0.22-μm filter plates and subjected to IL-1β ELISA analysis. A total of 200 μl of supernatant from 2 million isolated CD4 T cells was used for each condition. The assay was performed as described by the manufacturer's instructions (R&D Systems). Bioactive IL-1β was detected in supernatants of HIV-infected cultures, at levels comparable to those in uninfected cells treated with nigericin. Treatments of HIV-infected cultures with viral or caspase 1 inhibitors, but not caspase 3 inhibitor, reduced accumulation of IL-1β in the supernatants to levels comparable to those detected in uninfected cultures. These finding demonstrate that caspase 1 activation is specifically required for the release of bioactive IL-1β in lymphoid CD4 T cells infected with HIV-1. Error bars represent s.e.m. of three independent experiments using tonsil cells from at least three different donors. **b**, Inhibitors of caspase 1 and the NLRP3 inflammasome prevent release of mature IL-1β induced by nigericin, but not CD4 T-cell death by HIV-1. Because nigericin engages the NLRP3 inflammasome to activate caspase 1 in lymphoid CD4 T cells, we sought to determine if NLRP3 also similarly controls caspase 1 activity in response to HIV-1 infection. Cell cultures were treated with four separate NLRP3 inhibitors including CRID3[64], parthenolide[65], and the sulfonylureas glyburide[66] and glimepiride. Treatments with CRID3, parthenolide or sulfonylureas (not shown) completely inhibited NLRP3-dependent release of mature IL-1β by nigericin, but had no effect on IL-1β release triggered by HIV infection of lymphoid CD4 T-cell cultures (Fig. 3f). **c**, Treatments with CRID3, parthenolide or sulfonylureas did not prevent HIV-1-mediated CD4 T-cell death. These results suggest that the NLRP3 inflammasome does not control the caspase-1-mediated death responses in lymphoid CD4 T cells abortively infected with HIV-1. Cell death results are represented as ratios of viable CD4 versus CD8 T cells in each HIV-infected or uninfected culture. Error bars represent s.e.m. of four independent experiments using tonsil cells from four different donors.



Extended Data Figure 8 | Distinct regions of caspase 1 and caspase 3 activity in lymph node of a chronically infected HIV patient. Inguinal lymph node was collected from a 41-year-old African-American male, infected with an R5-tropic strain of HIV-1. The patient had been on intermittent anti-retroviral therapy between 2004–2009, and stopped anti-retroviral therapy in 2009. This individual exhibited a viral load of 30,173 RNA copies per ml, and CD4 T-cell count of 259 cells per µl. The fresh specimen was immediately subjected to immunostaining in parallel with fresh uninfected human tonsil. Note the immunostain against CD3 highlights the paracortical region, which is almost entirely composed of resting T cells. Note also the sparse presence of CD3-positive T cells in the mantle zones and germinal centres, where lymphocytes become activated (Ki-67) and differentiate into memory and plasma cells. These CD4 T cells are responsible for antigen-dependent activation of B cells in the follicle. Staining for CD11c reveals scattered dendritic cells[67,68] in the germinal centre and largely in the mantle zone. HIV p24 Gag expression is located between the mantle zone and germinal centres, where activated CD4 T cells reside. Remarkably, caspase 3 activity also occurs in this anatomical region, which is separated from the majority of non-productively infected T cells in the paracortical zone and exhibit caspase 1 activation, IL-1β processing and pyroptosis. The anti-caspase 1 antibody was raised against a peptide mapping to the C terminus of caspase 1 p20 of human origin and detects both the cleaved p20 subunit and the precursor of caspase 1. Therefore, in the absence of an equivalent uninfected lymph node it is hard to absolutely determine whether abortive HIV-1 infection affects pro-caspase expression. However, staining of uninfected tonsil or spleen (not shown) tissue revealed no positive HIV p24 Gag, active or pro-caspase 1, bioactive IL-1β or annexin V signals. These data closely correlate with the findings in HIV-infected HLACs where the 95% of the CD4 T cells are non-productively infected CD4 T cells and show activation of intracellular caspase 1, whereas caspase 3 activity is markedly less abundant and specifically occurs in productively infected CD4 T cells. GC, germinal centre; MZ, mantle zone; PC, paracortical zone.



Extended Data Figure 9 | Targeting caspase 1 via an orally bioavailable small molecule inhibitor prevents lymphoid CD4 T-cell death by HIV-1. **a**, VX-765 is a cell permeable pro-drug (**1**) that requires intracellular esterase cleavage in the cell to yield the aldehyde functionality (green) of the drug VRT-043298 (**2b**), which acts as a potent caspase 1 inhibitor. Adapted from ref. 38 with permission. **b**, VX-765 prevents CD4 T-cell death in a dose-dependent manner in HIV-infected lymphoid tissues. HLACs were either not infected or infected with HIV-1 in the absence of drugs or in the presence of efavirenz (100 nM), AMD3100 (250 nM) or VX-765 (0.05, 0.5 or 5 μM) as indicated. VX-765 was added to the cultures 4 hours before infection to allow absorption and processing by the cells. Flow cytometry plots depict gating on live cells based on the forward-scatter versus side-scatter profile of the complete culture. These results are representative of three independent experiments performed using tonsil cells from three different donors.



Extended Data Figure 10 | Treatment with a caspase1 inhibitor does not increase productive HIV-1 infection. To determine whether inhibition of caspase-1-mediated pyroptosis would result in higher levels of productive HIV-1 infection, tonsillar HLACs were treated with AMD3100 or with the caspase 1 inhibitor Ac-YVAD-CMK before infection with a GFP reporter virus (NLENG1). After 5 days, flow cytometry analysis of the infected cultures revealed no increase in GFP-positive cells in the infected cultures treated with the caspase 1 inhibitor Ac-YVAD-CMK. This result likely reflects the continued function of the host restriction factor SAMHD1 (refs 35, 69). These findings argue against the possibility that pyroptosis functions as a defence against productive infection. Instead, pyroptosis appears to represent an overall harmful response that centrally contributes to HIV pathogenesis. These results also argue that interdiction of the pyroptosis pathway with caspase 1 inhibitors would produce beneficial rather than harmful therapeutic effects.

ARTICLE

doi:10.1038/nature12890

Architecture of the large subunit of the mammalian mitochondrial ribosome

Basil J. Greber[1]*, Daniel Boehringer[1]*, Alexander Leitner[2], Philipp Bieri[1], Felix Voigts-Hoffmann[1], Jan P. Erzberger[1], Marc Leibundgut[1], Ruedi Aebersold[2,3] & Nenad Ban[1]

Mitochondrial ribosomes synthesize a number of highly hydrophobic proteins encoded on the genome of mitochondria, the organelles in eukaryotic cells that are responsible for energy conversion by oxidative phosphorylation. The ribosomes in mammalian mitochondria have undergone massive structural changes throughout their evolution, including ribosomal RNA shortening and acquisition of mitochondria-specific ribosomal proteins. Here we present the three-dimensional structure of the 39S large subunit of the porcine mitochondrial ribosome determined by cryo-electron microscopy at 4.9 Å resolution. The structure, combined with data from chemical crosslinking and mass spectrometry experiments, reveals the unique features of the 39S subunit at near-atomic resolution and provides detailed insight into the architecture of the polypeptide exit site. This region of the mitochondrial ribosome has been considerably remodelled compared to its bacterial counterpart, providing a specialized platform for the synthesis and membrane insertion of the highly hydrophobic protein components of the respiratory chain.

Mitochondrial ribosomes (mitoribosomes) are responsible for protein synthesis in mitochondria. These organelles of endosymbiotic origin[1] are required for energy conversion by aerobic respiration in eukaryotic cells. Mitoribosomes are more closely related to bacterial ribosomes than to eukaryotic cytosolic ribosomes[2]. However, the mammalian mitoribosome has been strongly altered by acquisition of mitochondria-specific ribosomal proteins and protein extensions[2–5], as well as the shortening of the mitochondrial ribosomal RNA (rRNA)[6]. The large 39S subunit of the mammalian mitoribosome catalyses peptide bond formation during protein synthesis and harbours the nascent polypeptide exit tunnel. The structural evolution of the mammalian mitoribosome was accompanied by a strong functional specialization towards the synthesis of the highly hydrophobic mitochondrial inner membrane proteins[7]. The region around the polypeptide tunnel exit of the mitoribosome serves as a specialized platform for membrane insertion and assembly of these critical mitochondrially encoded respiratory chain components[7–11]. Defects of the mitochondrial translation system are causally involved in a range of human diseases[12].

Although cryo-electron microscopy (cryo-EM) reconstructions of the bovine mitoribosome at 13.5 Å (ref. 13) and 12.1 Å (ref. 14) resolution have visualized large structural differences compared to the bacterial ribosome, detailed molecular insight into the architecture and arrangement of unique protein and rRNA elements of the mammalian mitoribosome is currently lacking. We have used cryo-EM combined with chemical crosslinking followed by mass spectrometry (CX-MS) to determine the structure of the large subunit of the mammalian mitoribosome, providing insight into its overall structure and into the molecular architecture of the polypeptide exit site in particular.

Structure determination

To obtain structural insights into the mammalian mitochondrial ribosome, we purified mitoribosomes and their subunits from porcine (*Sus scrofa*)

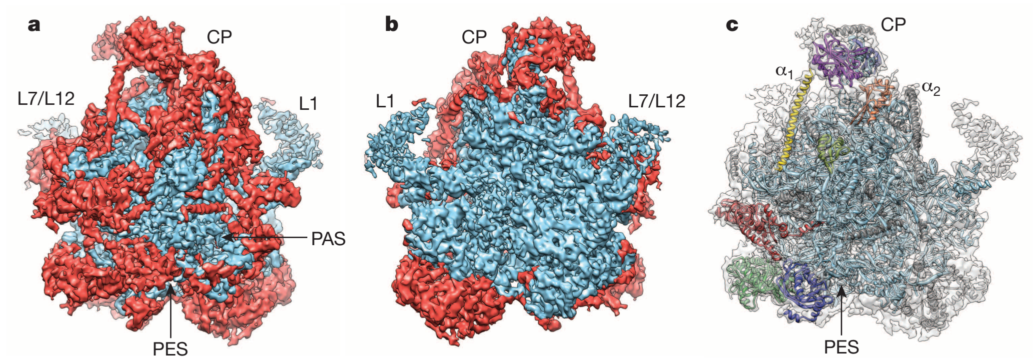


Figure 1 | Cryo-EM reconstruction of the mitoribosomal 39S subunit. **a**, **b**, Solvent side (**a**) and subunit interface side (**b**) of the 39S subunit segmented into conserved (blue) and mitoribosome-specific (red) density elements. CP, central protuberance; PAS, polypeptide accessible site; PES, polypeptide exit site; L1, L1 stalk; L7/L12, L7/L12 stalk base. **c**, Two clusters of mitochondria-specific ribosomal proteins (coloured) at the central protuberance and the polypeptide exit site. The central protuberance is flanked by prominent α-helices (α_1 and α_2). Conserved parts of the 39S subunit are in light blue; unassigned protein segments are in dark grey.

[1]Department of Biology, Institute of Molecular Biology and Biophysics, Schafmattstrasse 20, ETH Zurich, CH-8093 Zurich, Switzerland. [2]Department of Biology, Institute of Molecular Systems Biology, Wolfgang-Pauli-Strasse 16, ETH Zurich, CH-8093 Zurich, Switzerland. [3]Faculty of Science, University of Zurich, CH-8057 Zurich, Switzerland.
*These authors contributed equally to this work.

liver and collected cryo-EM data of the 39S mitoribosomal subunit. The cryo-EM reconstruction of the 39S ribosomal subunit (Fig. 1) at 4.9 Å resolution (Extended Data Fig. 1) extends the insights obtained from previous reconstructions[13], revealing the secondary structure elements of the mitoribosome-specific proteins and protein extensions that form a shell around the conserved core of ribosomal proteins and rRNA (Fig. 1a, b and Extended Data Fig. 2). The mitochondria-specific protein elements are mostly located on the solvent-exposed side of the 39S subunit, forming an extensive network of protein–protein contacts, whereas the subunit interface side is comparably better conserved (Fig. 1a, b), as observed previously in eukaryotic cytosolic ribosomes[15]. Because reliable *de novo* tracing of protein chains is not possible at the resolution of the cryo-EM map reported here, we subjected the *S. scrofa* 39S subunit and the *Bos taurus* 55S mitoribosome to CX-MS[16] to obtain protein–protein crosslinking data, enabling a molecular interpretation of our cryo-EM map (Fig. 1c).

Structure of the 16S rRNA

The density for the mitoribosomal 16S rRNA was clearly identifiable in the 4.9 Å cryo-EM map, with separation of individual strands and features that correspond to phosphate positions in well-ordered regions of the structure. Guided by the structure of the 23S rRNA of *Thermus thermophilus*[17] we were able to build a three-dimensional model of the strongly reduced 16S rRNA, in agreement with previous reports[13,14], but at a considerably higher level of detail (Fig. 2a, b and Extended Data Figs 3 and 4).

These reductions are most pronounced in domain I, where helices H7, H16, H18, H19 and H20 of the 23S rRNA are missing in the 16S rRNA. The resulting groove on the surface of the 39S subunit passes near the polypeptide tunnel exit, creating a lateral opening of the tunnel referred to as the polypeptide accessible site[13]. Additional losses of rRNA elements near the polypeptide tunnel exit occur in 16S rRNA domain III.

Although 16S rRNA is considered to be the only rRNA component of the 39S subunit of the mitoribosome[11], we observed additional RNA density at the central protuberance that cannot be accounted for by the 16S rRNA (Fig. 2c, d). This rRNA density is located in the immediate vicinity of MRPL18 and resembles the structural features of domain β of the bacterial 5S rRNA, which is located in the same area and interacts with the bacterial homologue of MRPL18 (Fig. 2d and Extended Data Fig. 5). However, the observed RNA density is insufficient to account for the entire 5S rRNA, and does not form a contact to the main body of the 39S subunit, as is the case for the 5S rRNA in bacterial and eukaryotic cytosolic ribosomes (Extended Data Fig. 5e, f). Taken together, our density clearly shows the presence of a second molecule of rRNA in the mammalian mitochondrial ribosome, possibly a portion of the 5S rRNA imported from the cytosol[18], or alternatively a different RNA species acting as a functional replacement for domain β of the 5S rRNA.

Novel proteins in the cryo-EM map of the 39S subunit

To analyse the protein component of the porcine 39S subunit, we first positioned homology models[19] of all mitoribosomal homologues of bacterial ribosomal proteins[2,5] visible in the map (Supplementary Table 1),

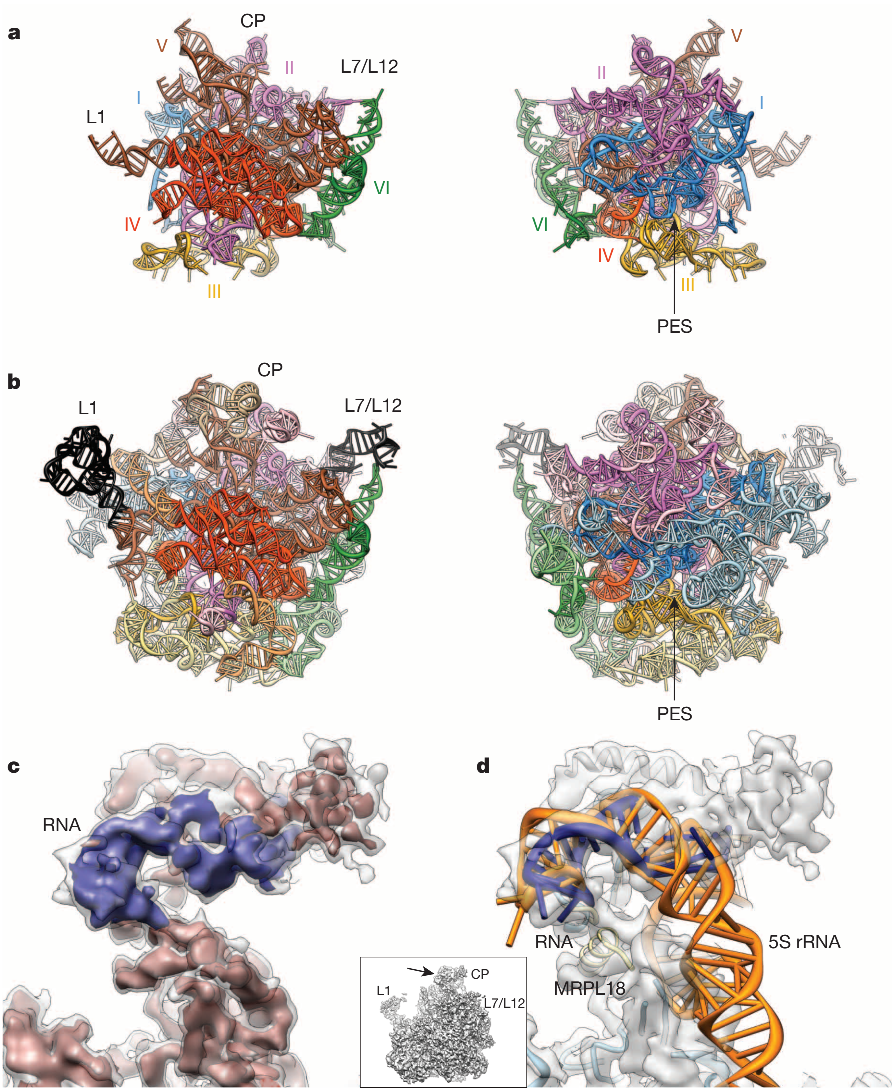


Figure 2 | The ribosomal RNA of the 39S mitoribosomal subunit. **a**, Domain architecture of the mitoribosomal 16S rRNA (left, subunit interface side; right, solvent side). Domains I–VI are colour coded, as indicated by Roman numerals. **b**, Bacterial 23S rRNA structure (Protein Data Bank (PDB) ID 3V2D)[17]. Colour code as in **a**, with segments absent in the mammalian mitoribosome shown in lighter colours. Black denotes segments not modelled in the 39S subunit owing to disorder. **c**, A clear rRNA major groove feature visualized in the cryo-EM map indicates the presence of a second rRNA molecule (blue) in the 39S subunit. **d**, Comparison of this mitochondrial rRNA (blue) and the bacterial 5S rRNA (orange, superposition according to MRPL18/L18).

leaving, however, considerable regions of density unexplained (Fig. 1). On the basis of visual inspection of the cryo-EM map and aided by the crosslinking data from our mass spectrometry experiments (Fig. 3 and Supplementary Table 2), we were able to position homology models of several mitochondria-specific proteins (Extended Data Fig. 6 and Supplementary Table 1). The mitochondria-specific ribosomal proteins MRPL39, MRPL44 and MRPL45 are localized at the polypeptide tunnel exit, and MRPL38, MRPL52, MRPL49 and ICT1 (also known as MRPL58) are located at the central protuberance.

Architecture of the central protuberance

The overall appearance of the mitoribosomal central protuberance is considerably different compared to its bacterial counterpart, featuring rod-like densities on each side[13] that can now be identified as α-helices. One of these helices emanates from the central protuberance in the immediate vicinity of the location we assigned to MRPL38 (Fig. 4a and Extended Data Fig. 6). Notably, our CX-MS data show the presence of several crosslinks of MRPL38 to MRPL52, a protein containing a very long predicted α-helix, thus establishing the identity of this helix as MRPL52 and confirming the assignment of MRPL38 (Fig. 4a, Extended Data Fig. 7 and Supplementary Table 3). MRPL52 therefore probably stabilizes the mitoribosomal central protuberance by forming contacts to both the body of the 39S subunit and MRPL38 at the top of the central protuberance, possibly compensating for the loss of the 5S rRNA contact to the subunit body (Extended Data Fig. 5e, f).

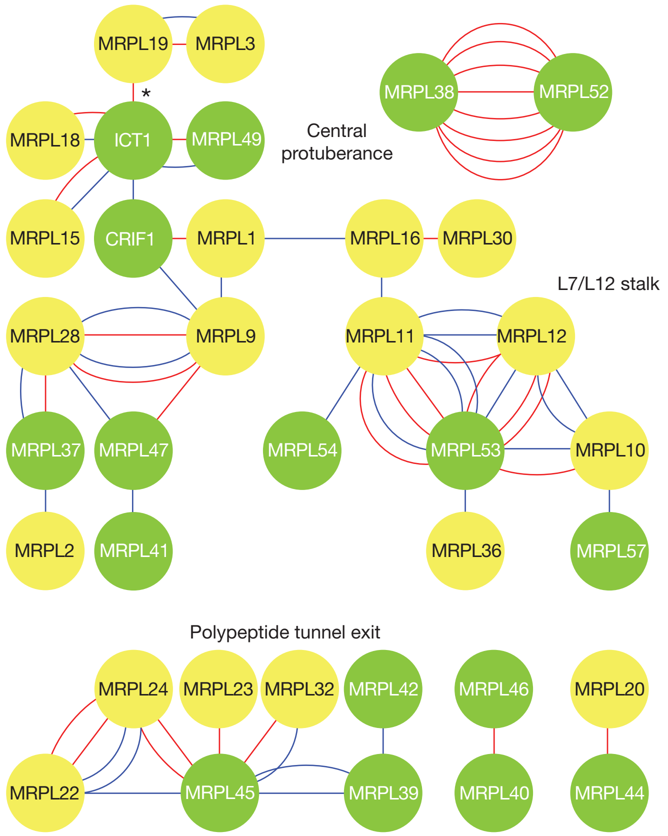


Figure 3 | Overview of the inter-protein crosslinks of 39S mitoribosomal subunit proteins. Statistically significant inter-protein crosslink hits of 39S subunit proteins (xQuest Id score >25) obtained by CX-MS experiments. Mitoribosomal proteins with bacterial ribosomal homologues are shown as yellow nodes, those without homologues in bacterial ribosomes in green. Crosslinks obtained using *S. scrofa* 39S subunits are shown as red lines, crosslinks obtained using *B. taurus* 55S mitoribosomes are shown in blue. The crosslink of ICT1 with MRPL19 (asterisk) in *S. scrofa* is likely a false positive identification (see Supplementary Table 3).

ICT1 is a functional peptidyl-transfer RNA hydrolase that has been stably incorporated into the mitoribosomal large subunit[20]. The structure of ICT1 (ref. 21) reveals that it is homologous to canonical bacterial class I release factors and the ribosomal rescue factor YaeJ, which bind to the ribosomal A-site[22,23]. In our CX-MS experiments, we observe several crosslinks between ICT1 and protein components of the mitoribosomal central protuberance (Fig. 3 and Supplementary Table 3). Furthermore, homology models of ICT1 and one of its crosslinking partners, MRPL49, can be fitted into the density next to each other, localizing ICT1 adjacent to MRPL38 and MRPL49 at the base of the central protuberance (Fig. 4b and Extended Data Fig. 6). This position of ICT1 differs from the positions of homologous domains of class I release factors in catalytically activated complexes, where they occupy the ribosomal A-site[22,23]. Consequently, ICT1 bound near MRPL38 cannot directly access the nascent chain to induce its release from a peptidyl-tRNA bound in the P-site. However, a permanent position in the A-site is not possible for ICT1 as this would impede translation. Therefore, ICT1 may have a purely architectural role in the 39S subunit, whereas its described activity in polypeptide release[20] may be performed by a pool of freely diffusing ICT1, or alternatively, polypeptide release by ICT1 may require large-scale displacements of either the P-site tRNA or ICT1 itself.

Architecture of the tunnel exit region

At the polypeptide exit site, density corresponding to the core folds of the mitoribosomal homologues of bacterial L23 (MRPL23), L29 (MRPL47 (ref. 5)), L22 (MRPL22), L24 (MRPL24) and L17 (MRPL17) indicates that the ring of proteins surrounding the polypeptide exit site is conserved (Extended Data Fig. 8). However, the tunnel exit region of the mammalian 39S subunit also shows a large number of mitochondria-specific features, including the previously described polypeptide accessible site[13], as well as a second layer of protein density, which we were able to partially interpret by docking homology models of the mitoribosome-specific proteins MRPL39, MRPL44 and MRPL45 (Fig. 4c, d and Extended Data Fig. 6). Their localization is consistent with the results of our CX-MS experiments (Extended Data Fig. 9 and Supplementary Table 3). MRPL39 and MRPL45 form a second layer of protein on top of the conserved proteins that surround the bacterial polypeptide tunnel exit. MRPL39 is located in close proximity of MRPL22, and MRPL45 is bound to MRPL22 and MRPL24 (Extended Data Fig. 8c). Both MRPL39 and MRPL44 associate with the 39S subunit via protein–protein interactions, but show homology to RNA-binding proteins, specifically to threonyl-tRNA synthetases[24], and to RNase III family members[4], respectively. Interestingly, the areas corresponding to RNA-binding surfaces in related proteins are oriented towards the solvent and not used for interactions with rRNA. The specific functions of MRPL39 and MRPL44, which is implicated in mitochondrial infantile cardiomyopathy (Extended Data Fig. 9c)[25], remain to be established.

A large fraction of mitoribosomes in mammalian mitochondria is tightly bound to the mitochondrial inner membrane independently of the presence of a nascent chain[26]. The position and orientation of MRPL45 next to the ribosomal tunnel exit and the structural features of this protein provide a possible explanation for this observation.

MRPL45 shows homology to the carboxy-terminal domain of TIM44 (ref. 27) and the *Saccharomyces cerevisiae* mitochondrial protein Mba1 (ref. 5), both of which are membrane-associated proteins[28,29]. Mba1 functions as a membrane-bound ribosome receptor in yeast[8,29] and has been crosslinked to ribosomal proteins in the vicinity of the ribosomal exit tunnel[9]. Furthermore, the orientation of MRPL45 on the 39S subunit reveals that the region that is structurally equivalent to the membrane-binding segment of TIM44 (ref. 30) is optimally positioned for membrane interactions (Extended Data Fig. 10). Taken together, these findings suggest that mammalian MRPL45 supports the binding of the mitoribosome to the mitochondrial inner membrane to align the nascent polypeptide tunnel exit with OXA1L (Oxa1 in yeast), which functions as an insertase for mitochondrial inner membrane proteins[7,31].

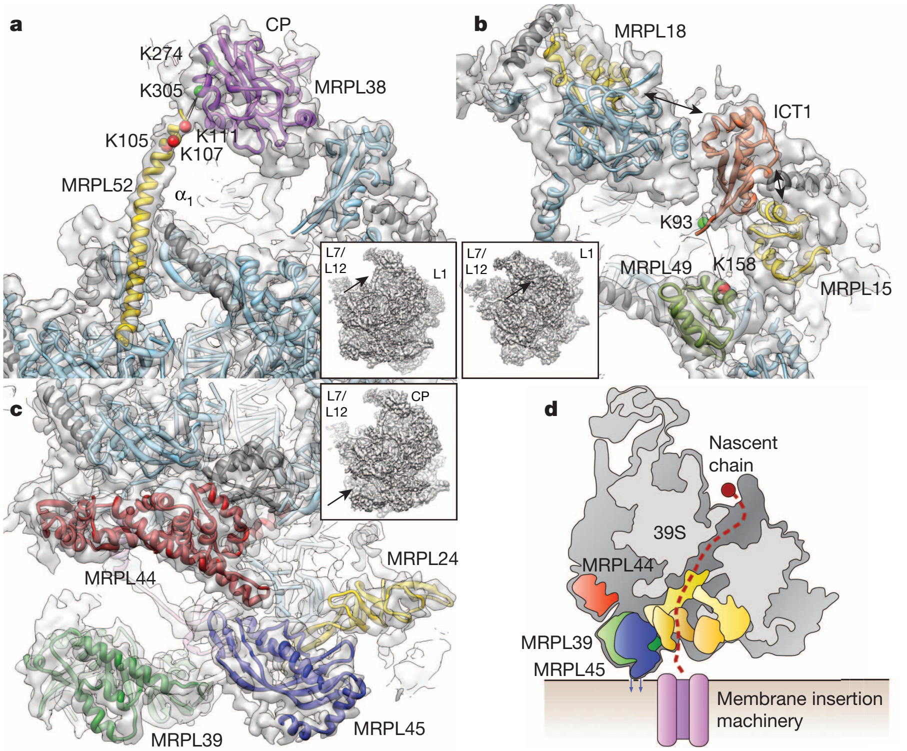


Figure 4 | Structures of novel mitochondria-specific proteins at the highly remodelled mitoribosomal central protuberance and at the polypeptide tunnel exit. **a**, Crosslinks between MRPL38 and MRPL52 allow assignment of MRPL52 to helical density α_1 (crosslinked residues shown as spheres if modelled, or as double arrows otherwise (**b**)). Insets show the whole 39S subunit for orientation. **b**, ICT1 (orange) and MRPL49 (olive) crosslink to each other and can be fitted into the cryo-EM map at the central protuberance. ICT1 also crosslinks to MRPL15 and MRPL18 (gold). **c**, Tunnel exit region of the 39S subunit with fitted MRPL39 (green), MRPL44 (red) and MRPL45 (blue). **d**, Schematic illustration of the 39S subunit bound to the mitochondrial inner membrane.

Although it was previously suggested that a large portion of the 39S subunit is inserted into the membrane[11], our data on the architecture of the 39S polypeptide exit site indicate that mammalian mitoribosomes are instead anchored to the membrane surface by at least one membrane-binding ribosomal protein, obviating the need for the universally conserved signal recognition particle membrane-targeting machinery[7].

Proteins functionally replace rRNA

Comparison of the 39S structure with structures of the prokaryotic ribosome[17,32] shows that the truncation of rRNA segments leads to the formation of deep channels on the surface of the subunit (Fig. 5a, b), some of which are partially occupied by novel mitoribosome-specific protein elements. Although smaller in size compared to the missing

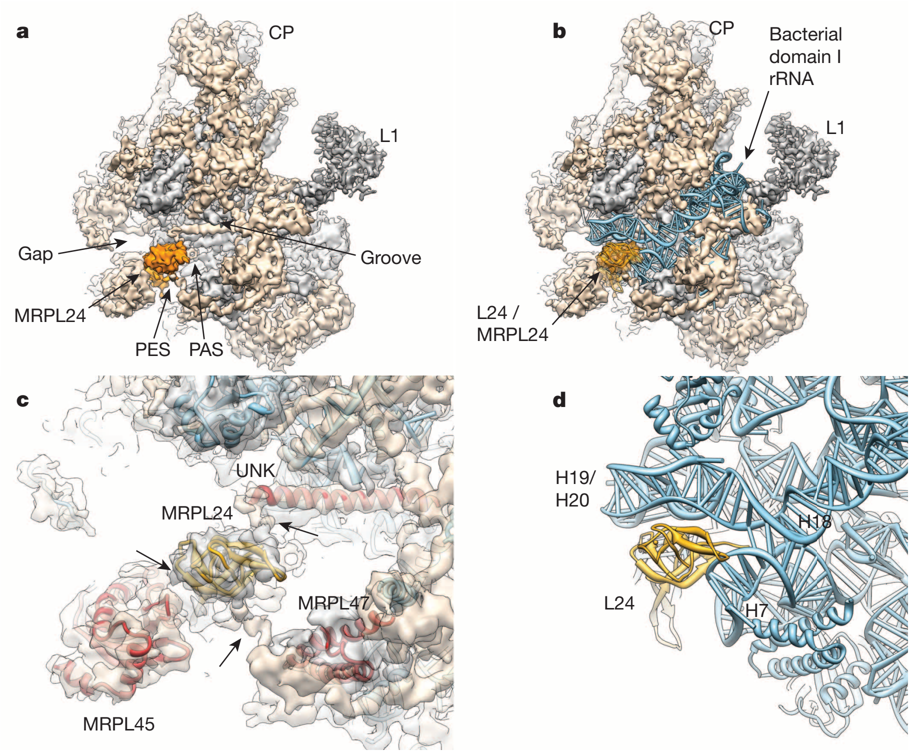


Figure 5 | Remodelling of MRPL24 contact sites in the 39S subunit. **a**, A deep groove runs above MRPL24, caused by extensive loss of rRNA compared to bacterial 23S rRNA (conserved parts grey; mitochondrial-specific parts brown; MRPL24 gold). **b**, 39S cryo-EM map shown together with a part of domain I of the bacterial 23S rRNA[17] (PDB ID 3V2D) (blue). **c**, Novel contacts (arrows) of mitoribosomal protein elements (red), among them MRPL45, additional density near MRPL47, and an unassigned α-helical density emanating from the core of the subunit (UNK), stabilize MRPL24. **d**, These mitochondrial-specific contacts of MRPL24 replace the extensive interactions of L24 with 23S rRNA in bacteria (colours as in **b**; labelled rRNA helices are missing in the 39S subunit).

bacterial rRNA segments, these proteins probably stabilize the mitoribosomal particle.

A particularly notable example of structural remodelling has occurred in the region near MRPL24 (Fig. 5). In the prokaryotic ribosome, ribosomal protein L24 forms extensive contacts to the 23S rRNA helices H7 and H19 (Fig. 5b, d)[17,32]. These helices have been lost in the mammalian mitoribosome, eliminating almost all bacterial-like contacts of MRPL24 in the 39S subunit (Fig. 5a, c). However, the mitoribosome-specific protein MRPL45 and two currently unassigned additional protein elements form novel contacts with MRPL24, holding it in place with minimal change in its position and orientation compared to the bacterial ribosome (Fig. 5c). Even though these proteins do not directly replace the missing rRNA structure, they act as an architectural replacement to maintain the position of the functionally important protein MRPL24 (ref. 7) relative to the ribosomal tunnel. Owing to the nearly complete remodelling of contacts, this process must have occurred in a stepwise fashion during evolution.

In conclusion, the mitoribosome has undergone a reduction in rRNA content, which is architecturally compensated for by mitoribosome-specific proteins, some of which bear structural and functional homology to soluble and membrane-associated proteins or enzymes. The tunnel exit region is considerably remodelled, which is probably an adaptation of the 39S subunit to the highly specific requirements of the synthesis and membrane insertion of respiratory chain proteins. Our identification of MRPL45 at the mitoribosomal tunnel exit provides a structural explanation for the observed membrane attachment of mitochondrial ribosomes (Fig. 4d). The results reported here provide an excellent starting point for functional studies.

METHODS SUMMARY

55S mitoribosomes and 39S mitoribosomal subunits were prepared from porcine (*Sus scrofa*) and bovine (*Bos taurus*) mitochondria obtained from liver tissue. The three-dimensional structures of the 55S mitoribosome and the 39S mitoribosomal subunit were determined by cryo-EM. Chemical crosslinking followed by mass spectrometry was used to obtain protein–protein crosslinking data. Proteins and rRNA were built into the cryo-EM density on the basis of homology modelling and crosslinking data.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 September; accepted 15 November 2013.
Published online 22 December 2013.

1. Sagan, L. On the origin of mitosing cells. *J. Theor. Biol.* **14,** 225–274 (1967).
2. Desmond, E., Brochier-Armanet, C., Forterre, P. & Gribaldo, S. On the last common ancestor and early evolution of eukaryotes: reconstructing the history of mitochondrial ribosomes. *Res. Microbiol.* **162,** 53–70 (2011).
3. Suzuki, T. *et al.* Structural compensation for the deficit of rRNA with proteins in the mammalian mitochondrial ribosome. Systematic analysis of protein components of the large ribosomal subunit from mammalian mitochondria. *J. Biol. Chem.* **276,** 21724–21736 (2001).
4. Koc, E. C. *et al.* The large subunit of the mammalian mitochondrial ribosome. Analysis of the complement of ribosomal proteins present. *J. Biol. Chem.* **276,** 43958–43969 (2001).
5. Smits, P., Smeitink, J. A. M., van den Heuvel, L. P., Huynen, M. A. & Ettema, T. J. G. Reconstructing the evolution of the mitochondrial ribosomal proteome. *Nucleic Acids Res.* **35,** 4686–4703 (2007).
6. Attardi, G. & Ojala, D. Mitochondrial ribosome in HeLa cells. *Nat. New Biol.* **229,** 133–136 (1971).
7. Ott, M. & Herrmann, J. M. Co-translational membrane insertion of mitochondrially encoded proteins. *Biochim. Biophys. Acta* **1803,** 767–775 (2010).
8. Ott, M. *et al.* Mba1, a membrane-associated ribosome receptor in mitochondria. *EMBO J.* **25,** 1603–1610 (2006).
9. Gruschke, S. *et al.* Proteins at the polypeptide tunnel exit of the yeast mitochondrial ribosome. *J. Biol. Chem.* **285,** 19022–19028 (2010).
10. Jia, L. *et al.* Yeast Oxa1 interacts with mitochondrial ribosomes: the importance of the C-terminal region of Oxa1. *EMBO J.* **22,** 6438–6447 (2003).
11. Agrawal, R. K. & Sharma, M. R. Structural aspects of mitochondrial translational apparatus. *Curr. Opin. Struct. Biol.* **22,** 797–803 (2012).
12. Pearce, S., Nezich, C. L. & Spinazzola, A. Mitochondrial diseases: translation matters. *Mol. Cell. Neurosci.* **55,** 1–12 (2013).
13. Sharma, M. R. *et al.* Structure of the mammalian mitochondrial ribosome reveals an expanded functional role for its component proteins. *Cell* **115,** 97–108 (2003).
14. Mears, J. A. *et al.* A structural model for the large subunit of the mammalian mitochondrial ribosome. *J. Mol. Biol.* **358,** 193–212 (2006).
15. Klinge, S., Voigts-Hoffmann, F., Leibundgut, M. & Ban, N. Atomic structures of the eukaryotic ribosome. *Trends Biochem. Sci.* **37,** 189–198 (2012).
16. Walzthoeni, T., Leitner, A., Stengel, F. & Aebersold, R. Mass spectrometry supported determination of protein complex structure. *Curr. Opin. Struct. Biol.* **23,** 252–260 (2013).
17. Polikanov, Y. S., Blaha, G. M. & Steitz, T. A. How hibernation factors RMF, HPF, and YfiA turn off protein synthesis. *Science* **336,** 915–918 (2012).
18. Smirnov, A., Entelis, N., Martin, R. P. & Tarassov, I. Biological significance of 5S rRNA import into human mitochondria: role of ribosomal protein MRP-L18. *Genes Dev.* **25,** 1289–1305 (2011).
19. Kelley, L. A. & Sternberg, M. J. E. Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* **4,** 363–371 (2009).
20. Richter, R. *et al.* A functional peptidyl-tRNA hydrolase, ICT1, has been recruited into the human mitochondrial ribosome. *EMBO J.* **29,** 1116–1125 (2010).
21. Handa, Y. *et al.* Solution structure of the catalytic domain of the mitochondrial protein ICT1 that is essential for cell vitality. *J. Mol. Biol.* **404,** 260–273 (2010).
22. Laurberg, M. *et al.* Structural basis for translation termination on the 70S ribosome. *Nature* **454,** 852–857 (2008).
23. Gagnon, M. G., Seetharaman, S. V., Bulkley, D. & Steitz, T. A. Structural basis for the rescue of stalled ribosomes: structure of YaeJ bound to the ribosome. *Science* **335,** 1370–1372 (2012).
24. Spirina, O. *et al.* Heart-specific splice-variant of a human mitochondrial ribosomal protein (mRNA processing; tissue specific splicing). *Gene* **261,** 229–234 (2000).
25. Carroll, C. J. *et al.* Whole-exome sequencing identifies a mutation in the mitochondrial ribosome protein MRPL44 to underlie mitochondrial infantile cardiomyopathy. *J. Med. Genet.* **50,** 151–159 (2013).
26. Liu, M. & Spremulli, L. Interaction of mammalian mitochondrial ribosomes with the inner membrane. *J. Biol. Chem.* **275,** 29400–29406 (2000).
27. Schneider, H. C. *et al.* Mitochondrial Hsp70/MIM44 complex facilitates protein import. *Nature* **371,** 768–774 (1994).
28. Weiss, C. *et al.* Domain structure and lipid interaction of recombinant yeast Tim44. *Proc. Natl Acad. Sci. USA* **96,** 8890–8894 (1999).
29. Preuss, M. *et al.* Mba1, a novel component of the mitochondrial protein export machinery of the yeast *Saccharomyces cerevisiae. J. Cell Biol.* **153,** 1085–1096 (2001).
30. Cui, W., Josyula, R., Li, J., Fu, Z. & Sha, B. Membrane binding mechanism of yeast mitochondrial peripheral membrane protein TIM44. *Protein Pept. Lett.* **18,** 718–725 (2011).
31. Stiburek, L. *et al.* Knockdown of human Oxa1l impairs the biogenesis of F_1F_o-ATP synthase and NADH:ubiquinone oxidoreductase. *J. Mol. Biol.* **374,** 506–516 (2007).
32. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289,** 905–920 (2000).

Supplementary Information is available in the online version of the paper.

Acknowledgements Cryo-EM data were collected at the electron microscopy facility of ETH Zurich (EMEZ) and at FEI Company Eindhoven. We thank P. Tittmann, F. de Haas and K. Sader for support. We acknowledge the use of computing infrastructure provided by the Central Information Technology Services of ETH Zurich. This work was supported by the Swiss National Science Foundation (SNSF), the National Center of Excellence in Research (NCCR) Structural Biology program of the SNSF, European Research Council (ERC) grant 250071 under the European Community's Seventh Framework Programme (to N.B.), the Commission of the European Communities through the PROSPECTS consortium (EU FP7 projects 201648, 233226) (R.A.) and the European Research Council (ERC-2008-AdG 233226) (R.A.).

Author Contributions F.V.-H., J.P.E. and N.B. initiated the project; F.V.-H. and J.P.E. established the purification procedures. F.V.-H., B.J.G. and P.B. performed preparation of the mitoribosomes. B.J.G., P.B. and D.B. prepared cryo-EM samples. D.B. acquired the cryo-EM data. B.J.G., D.B. and P.B. calculated the cryo-EM reconstructions. M.L., B.J.G., P.B., D.B. and N.B. interpreted the structure. A.L. performed CX-MS experiments in the laboratory of R.A. All authors contributed to the final version of the paper.

Author Information The cryo-EM map of the 39S mitoribosomal subunit has been deposited in the Electron Microscopy Databank with accession code EMD-2490. The coordinates of the cryo-EM-based model of the 39S mitoribosomal subunit have been deposited in the Protein Data Bank under accession code 4CE4. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.B. (ban@mol.biol.ethz.ch) or R.A. (aebersold@imsb.biol.ethz.ch).

doi:10.1038/nature12917

A millisecond pulsar in a stellar triple system

S. M. Ransom[1], I. H. Stairs[2], A. M. Archibald[3,4], J. W. T. Hessels[3,5], D. L. Kaplan[6,7], M. H. van Kerkwijk[8], J. Boyles[9,10], A. T. Deller[3], S. Chatterjee[11], A. Schechtman-Rook[7], A. Berndsen[2], R. S. Lynch[4], D. R. Lorimer[9], C. Karako-Argaman[4], V. M. Kaspi[4], V. I. Kondratiev[3,12], M. A. McLaughlin[9], J. van Leeuwen[3,5], R. Rosen[1,9], M. S. E. Roberts[13,14] & K. Stovall[15,16]

Gravitationally bound three-body systems have been studied for hundreds of years[1,2] and are common in our Galaxy[3,4]. They show complex orbital interactions, which can constrain the compositions, masses and interior structures of the bodies[5] and test theories of gravity[6], if sufficiently precise measurements are available. A triple system containing a radio pulsar could provide such measurements, but the only previously known such system, PSR B1620-26 (refs 7, 8; with a millisecond pulsar, a white dwarf, and a planetary-mass object in an orbit of several decades), shows only weak interactions. Here we report precision timing and multiwavelength observations of PSR J0337+1715, a millisecond pulsar in a hierarchical triple system with two other stars. Strong gravitational interactions are apparent and provide the masses of the pulsar (1.4378(13)$M_\odot$, where $M_\odot$ is the solar mass and the parentheses contain the uncertainty in the final decimal places) and the two white dwarf companions (0.19751(15)$M_\odot$ and 0.4101(3)$M_\odot$), as well as the inclinations of the orbits (both about 39.2°). The unexpectedly coplanar and nearly circular orbits indicate a complex and exotic evolutionary past that differs from those of known stellar systems. The gravitational field of the outer white dwarf strongly accelerates the inner binary containing the neutron star, and the system will thus provide an ideal laboratory in which to test the strong equivalence principle of general relativity.

Millisecond pulsars (MSPs) are neutron stars that rotate hundreds of times per second and emit beams of radio waves much as a lighthouse emits light. They are thought to form in binary systems[9] and their rotation rates and orbital properties can be measured with extremely high precision using the unambiguous pulse-counting methodology known as pulsar timing. As part of a large-scale pulsar survey[10,11] with the Robert C. Byrd Green Bank Telescope (GBT), we have discovered the only known MSP in a stellar triple system. The pulsar has a spin period of 2.73 ms, is relatively bright (~2 mJy for emission at 1.4 GHz), and has a complex radio pulse profile with multiple narrow components.

Although initial timing observations showed a seemingly typical binary MSP system with a 1.6-day circular orbit and a 0.1$M_\odot$–0.2$M_\odot$ white dwarf companion, large timing systematics quickly appeared, strongly suggesting the presence of a third body. Two other MSPs are known to have multiple companions: the pulsar B1257+12, which hosts at least three low-mass planets[12,13], and the MSP triple system B1620−26 in globular cluster M4, which has a white dwarf inner companion and an outer companion of roughly the mass of Jupiter[7,8]. The timing perturbations from J0337+1715 were much too large to be caused by a planetary-mass companion.

We began an intensive multifrequency radio timing campaign (Methods) using the GBT, the Arecibo telescope and the Westerbork Synthesis Radio Telescope (WSRT) to constrain the system's position and orbital parameters and the nature of the third body. At Arecibo, we achieved median arrival time uncertainties of 0.8 μs in 10 s, implying that half-hour integrations provide a precision of ~100 ns, making J0337+1715 one of the MSPs with highest known timing precisions.

To fold the pulsar signal (that is, to sum the data modulo the observed pulsar spin period), we approximate the motion of J0337+1715 using a pair of Keplerian orbits, with the centre of mass of the inner orbit moving around in the outer orbit. We determine pulse times of arrival (TOAs) from the folded radio data using standard techniques (Methods) and then correct them to the Solar System barycentre at infinite frequency using a precise radio position obtained with the Very Long Baseline Array (VLBA; Methods). These TOAs vary significantly, by the Rømer and Einstein delays[14], from those predicted by a simple pulsar spin-down model. The Rømer delay is a simple geometric effect due to the finite speed of light, and measures the pulsar's orbital motion. Its amplitude is $a_I\sin(i)/c \approx 1.2$ s (where a_I is semimajor axis of the inner orbit, i is the inclination of the orbit and c is the speed of light) for the inner orbit and ~74.6 s for the outer orbit (Figs 1 and 2).

The Einstein delay is the cumulative effect of time dilation—both special relativistic, due to the transverse Doppler effect, and general relativistic, due to gravitational redshift resulting from the pulsar's position in the total gravitational potential of the system. For J0337+1715, the gravitational redshift portion is covariant with fitting the projected semimajor axis of the orbit, just as the full Einstein delay is for other pulsars with circular orbits. The transverse Doppler effect is easily measurable, though, because it is proportional to $v^2/c^2 = |v_I + v_O|^2/c^2 = (|v_I|^2 + |v_O|^2 + 2v_I\bullet v_O)/c^2$ where v_I and v_O are the three-dimensional velocities in the inner and outer orbits, respectively. The squared magnitudes are covariant with orbital fitting as in the binary case, but the cross term, $v_I\bullet v_O/c^2$, contributes delays of tens of microseconds on the timescale of the inner orbit.

The two-Keplerian-orbit approximation results in systematic errors of up to several hundred microseconds over multiple timescales, owing to three-body interactions not accounted for by the model (Fig. 1), but those systematics carry a great deal of information about the system masses and geometry. The planet pulsar B1257+12 showed similar systematics[12], and direct numerical integrations confirmed their planetary nature and provided the masses and orbits of the planets[13,15]. The interactions in J0337+1715 are many orders of magnitude greater, but they, along with the Rømer and Einstein delays, can be similarly modelled by direct integration.

We use Monte Carlo techniques (Methods) to find sets of parameter values that minimize the difference between measured TOAs and TOAs predicted by three-body integrations, and we determine their

[1]National Radio Astronomy Observatory, 520 Edgemont Road, Charlottesville, Virginia 22903-2475, USA. [2]Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, British Columbia V6T 1Z1, Canada. [3]Netherlands Institute for Radio Astronomy (ASTRON), Postbus 2, 7990 AA Dwingeloo, The Netherlands. [4]Department of Physics, McGill University, 3600 University Street, Montreal, Quebec H3A 2T8, Canada. [5]Astronomical Institute 'Anton Pannekoek', University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands. [6]Physics Department, University of Wisconsin-Milwaukee, PO Box 413, Milwaukee, Wisconsin 53201, USA. [7]Department of Astronomy, University of Wisconsin-Madison, 475 North Charter Street, Madison, Wisconsin 53706-1582, USA. [8]Department of Astronomy and Astrophysics, University of Toronto, 50 St George Street, Toronto, Ontario M5S 3H4, Canada. [9]Department of Physics and Astronomy, West Virginia University, White Hall, Box 6315, Morgantown, West Virginia 26506-6315, USA. [10]Physics and Astronomy Department, Western Kentucky University, 1906 College Heights Boulevard, Bowling Green, Kentucky 42101-1077, USA. [11]Center for Radiophysics and Space Research, Cornell University, 524 Space Sciences Building, Ithaca, New York 14853, USA. [12]Astro Space Center of the Lebedev Physical Institute, 53 Leninskij Prospekt, Moscow 119991, Russia. [13]Eureka Scientific Inc., 2452 Delmer Street, Suite 100, Oakland, California 94602-3017, USA. [14]Physics Department, New York University at Abu Dhabi, PO Box 129188, Abu Dhabi, United Arab Emirates. [15]Department of Physics and Astronomy, University of Texas at Brownsville, One West University Boulevard, Brownsville, Texas 78520, USA. [16]Physics and Astronomy Department, University of New Mexico, 1919 Lomas Boulevard NE, Albuquerque, New Mexico 87131-0001, USA.

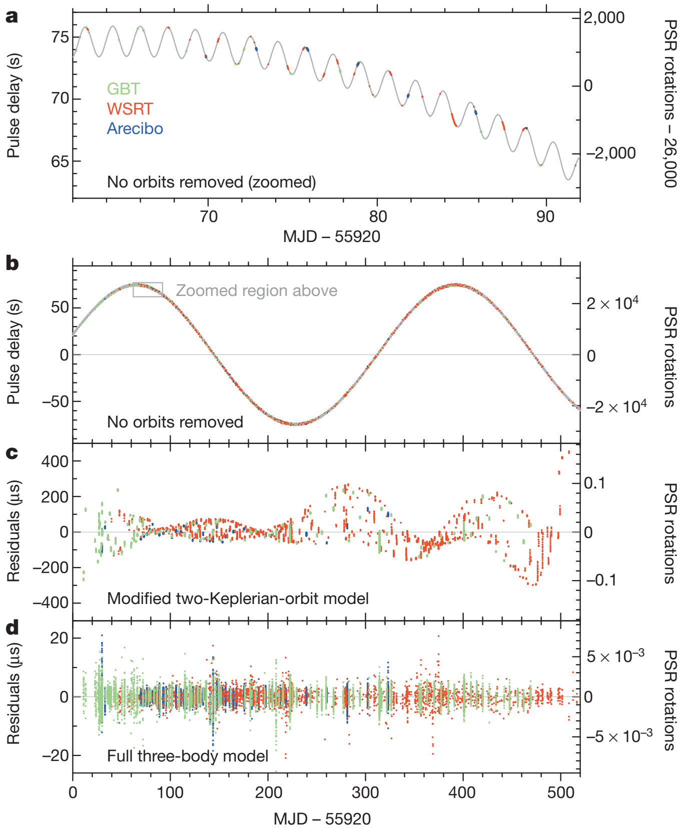


Figure 1 | Timing residuals and delays from the PSR J0337+1715 system. **a**, **b**, Geometric light-travel time delays (that is, Rømer delays), in both time and pulse periods, across the inner (**a**) and outer (**b**) orbits, and modified Julian dates (MJD) of radio timing observations from the GBT, the WSRT and the Arecibo telescope. Arrival time errors in these panels are approximately a million times too small to see. **c**, Newtonian three-body perturbations compared with the modified two-Keplerian-orbit model used for folding our data at the observed pulse period. **d**, Post-fit timing residuals from our full Markov chain Monte Carlo (MCMC)-derived three-body timing solution described in Table 1. The weighted root mean squared value of the 26,280 residuals is 1.34 μs.

expected values and error estimates directly from the parameter posterior distributions. We plot the results in Fig. 1 and list best-fit parameters and several derived quantities in Table 1. Component masses and relative inclinations are determined at the 0.1%–0.01% level, which is one to two orders of magnitude more precisely than from other MSP timing experiments, by a method that is effectively independent of the gravitational theory used. A detailed description of the three-body model and fitting procedure is under way (A.M.A. *et al.*, manuscript in preparation).

Using an early radio position, we identified an object with unusually blue colours in the Sloan Digital Sky Survey[16] (SDSS; Fig. 3). The optical and archival ultraviolet photometry, combined with new near- and mid-infrared photometry, are consistent (Methods) with a single white dwarf of temperature ~15,000 K, which optical spectroscopy confirmed is the inner white dwarf in the system (D.L.K. *et al.*, manuscript in preparation). When combined with the known white dwarf mass from timing observations, white dwarf models provide a radius from which we infer a photometric distance to the system of 1,300 ± 80 pc. The photometry and timing masses also exclude the possibility that the outer companion is a main-sequence star.

The pulsar in this system seems to be a typical radio MSP, but it is unique in having two white dwarf companions in hierarchical orbits. Although more than 300 MSPs are known in the Galaxy and in globular clusters, J0337+1715 is the first MSP stellar triple system found. Because there are no significant observational selection effects discriminating against the discovery of pulsar triple (as opposed to binary) systems, this implies that $\lesssim$1% of the MSP population resides in stellar triples and that $\lesssim$100 such systems exist in the Galaxy.

Predictions for the population of MSP stellar triples have suggested that most have highly eccentric outer orbits owing to dynamical interactions between the stars during stellar evolution[17]. Such models could also produce eccentric binaries such as MSP J1903+0327 (ref. 18), if the inner white dwarf, which had previously recycled the pulsar (that is, turned it into an MSP through the transfer of matter and angular momentum), were destroyed or ejected from the system dynamically[19]. In such situations, however, the coplanarity and circularity of the orbits of J0337+1715 would be very surprising. Those orbital characteristics, and their highly hierarchical nature ($P_{b,O}/P_{b,I} \approx 200$, where $P_{b,O}$ and $P_{b,I}$ are the orbital periods for the outer and inner binaries, respectively), imply that the current configuration is stable on long timescales[20], greatly increasing the odds of observing a triple system such as J0337+1715. Secular changes to the various orbital parameters will

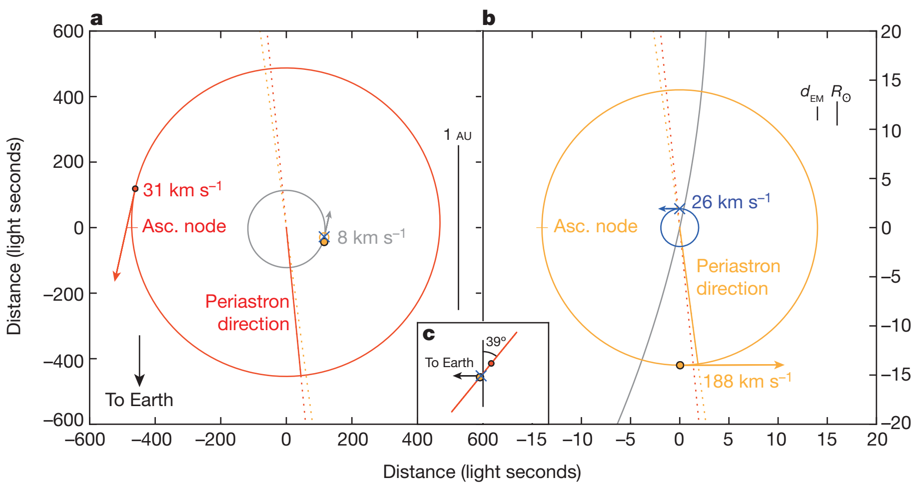


Figure 2 | Geometry of the PSR J0337+1715 system at the reference epoch. **a**, Orbital shape and velocity of the outer white dwarf (red), and the orbital shape and velocity of the centre of mass of the inner binary (grey). **b**, Orbital shapes and velocities of the inner white dwarf (orange) and the pulsar (blue). Dotted red and orange lines indicate the directions of periastron for the inner and outer white dwarf orbits, respectively. The white dwarf positions when the pulsar or inner orbit centre of mass crosses the ascending nodes are indicated. Vertical lines show length scales in the system in astronomical units (AU; **a**) or the Earth–Moon distance (d_{EM}) and the Solar radius ($R_\odot$; **b**). **c**, Inclination of the basically coplanar orbits with respect to the Earth–pulsar direction.

Table 1 | System parameters for PSR J0337+1715

Parameter	Value
Fixed values	
Right ascension, RA	03 h 37 min 43.82589(13) s
Declination, dec.	$17^{\circ}\ 15'\ 14.828(2)''$
Dispersion measure, DM	21.3162(3) pc cm^{-3}
Solar System ephemeris	DE405
Reference epoch	MJD 55920.0
Observation span	MJD 55930.9–56436.5
Number of TOAs	26,280
Weighted root mean squared residual	1.34 μs
Fitted parameters	
Spin-down parameters	
Pulsar spin frequency, f	365.953363096(11) Hz
Spin frequency derivative, $\dot{f}$	$-2.3658(12) \times 10^{-15}$ Hz s^{-1}
Inner Keplerian parameters for pulsar orbit	
Semimajor axis projected along line of sight, $(a\sin i)_{\mathrm{I}}$	1.21752844(4) ls
Orbital period, $P_{\mathrm{b,I}}$	1.629401788(5) d
Eccentricity parameter, $\varepsilon_{1,I} = (e\sin\omega)_{\mathrm{I}}$	$6.8567(2) \times 10^{-4}$
Eccentricity parameter, $\varepsilon_{2,I} = (e\cos\omega)_{\mathrm{I}}$	$-9.171(2) \times 10^{-5}$
Time of ascending node, $t_{\mathrm{asc,I}}$	MJD 55920.407717436(17)
Outer Keplerian parameters for centre of mass of inner binary	
Semimajor axis projected along line of sight, $(a\sin i)_{\mathrm{O}}$	74.6727101(8) ls
Orbital period, $P_{\mathrm{b,O}}$	327.257541(7) d
Eccentricity parameter, $\varepsilon_{1,\mathrm{O}} = (e\sin\omega)_{\mathrm{O}}$	$3.5186279(3) \times 10^{-2}$
Eccentricity parameter, $\varepsilon_{2,\mathrm{O}} = (e\cos\omega)_{\mathrm{O}}$	$-3.462131(11) \times 10^{-3}$
Time of ascending node, $t_{\mathrm{asc,O}}$	MJD 56233.935815(7)
Interaction parameters	
Semimajor axis projected in plane of sky, $(a\cos i)_{\mathrm{I}}$	1.4900(5) ls
Semimajor axis projected in plane of sky, $(a\cos i)_{\mathrm{O}}$	91.42(4) ls
Ratio of inner companion mass to pulsar mass, $q_{\mathrm{I}} = m_{\mathrm{cI}}/m_{\mathrm{p}}$	0.13737(4)
Difference in longitudes of asc. nodes, δ_{Ω}	$2.7(6) \times 10^{-3\circ}$
Inferred or derived values	
Pulsar properties	
Pulsar period, P	2.73258863244(9) ms
Pulsar period derivative, $\dot{P}$	$1.7666(9) \times 10^{-20}$
Inferred surface dipole magnetic field, B	2.2×10^{8} G
Spin-down power, $\dot{E}$	3.4×10^{34} erg s^{-1}
Characteristic age, τ	2.5×10^{9} yr
Orbital geometry	
Pulsar semimajor axis (inner), a_{I}	1.9242(4) ls
Eccentricity (inner), e_{I}	$6.9178(2) \times 10^{-4}$
Longitude of periastron (inner), ω_{I}	97.6182(19)°
Pulsar semimajor axis (outer), a_{O}	118.04(3) ls
Eccentricity (outer), e_{O}	$3.53561955(17) \times 10^{-2}$
Longitude of periastron (outer), ω_{O}	95.619493(19)°
Inclination of invariant plane, i	39.243(11)°
Inclination of inner orbit, i_{I}	39.254(10)°
Angle between orbital planes, δ_i	$1.20(17) \times 10^{-2\circ}$
Angle between eccentricity vectors, $\delta_{\omega} \approx \omega_O - \omega_I$	−1.9987(19)°
Masses	
Pulsar mass, m_{p}	$1.4378(13)M_{\odot}$
Inner companion mass, m_{cI}	$0.19751(15)M_{\odot}$
Outer companion mass, m_{cO}	$0.4101(3)M_{\odot}$

Values in parentheses represent 1σ errors in the final decimal place(s), as determined by MCMC fitting (Methods). Fixed values are parameters supplied as input to our timing fit; the position was obtained from an observation with the VLBA, and the DM was measured from high-signal-to-noise Arecibo telescope observations. Fitted parameters are those used to describe the state of the system at the reference epoch—the initial conditions for the differential equation integrator. Along with the pulsar spin-down parameters, these parameters include the conventional Keplerian elements measurable in binary pulsars for each of the orbits, plus four parameters measurable only owing to three-body interactions. These 14 orbital parameters can completely describe any configuration of three masses, positions and velocities for which the centre of mass remains fixed at the origin, provided that the longitude of the inner ascending node is zero. Although some fitting parameters are highly covariant, we computed all parameters and their errors on the basis of the posterior distributions, taking into account these covariances. We use the standard formulae for computing B, $\dot{E}$ and τ, which assume a pulsar mass of $1.4M_{\odot}$ and a moment of inertia of 10^{45} g cm^2. We have not corrected these values for proper motion or Galactic acceleration. The Laplace–Lagrange parameters, ε_1 and ε_2, parameterize eccentric orbits in a way that avoids a coordinate singularity at zero eccentricity. The pair $(\varepsilon_1, \varepsilon_2)$ forms a vector in the plane of the orbit called the eccentricity vector. For a single orbit, the ascending node is the place where the pulsar passes through the plane of the sky moving away from Earth; the longitude of the ascending node specifies the orientation of the orbit on the sky. This is not measurable with the data we have, but the difference between the longitudes of the ascending nodes of the two orbits is measurable from orbital interactions. The invariant plane is the plane perpendicular to the total (orbital) angular momentum of the triple system. ls, light second.

occur in the long term[21], however, and the three-body integrations and timing observations will predict and measure them.

The basic evolution of the system, which was almost certainly complex and exotic, may have progressed as follows. The most massive of the progenitor stars evolved off the main sequence and exploded in a supernova, creating the neutron star. At least two of the companions to the original primary survived the explosion, probably in eccentric orbits. After a period of order 10^9 yr, the outermost star, the next most massive, evolved and transferred mass onto the inner binary, which comprised the neutron star and a lower-mass main-sequence star, perhaps within a common envelope. During this phase, the angular momentum vectors of the inner and outer orbits were torqued into near alignment[22]. After the outer star ejected its envelope to become a white dwarf and another period of order 10^9 yr passed, the remaining main-sequence star evolved and recycled the neutron star as in the standard scenario[23]. During this phase, the inner orbit became highly circular but only a small amount of mass ($<0.2M_{\odot}$ in total) was transferred to the neutron star—enough to increase its rotation rate drastically, but not enough to make it particularly massive[24,25]. Since then, secular effects due to three-body interactions have aligned the apsides of the two orbits[5], although our three-body integrations display librations around apsidal alignment on both outer-orbital and secular timescales. This scenario explains well the coplanarity and circularity of the orbits as well as the fact that both white dwarf companions fall on the predicted curve relating helium white dwarf mass and orbital period[23].

Perhaps the most interesting aspect of J0337+1715 is its potential to provide extremely sensitive tests of the strong equivalence principle[26,27] (SEP), a key implication of which is that the orbital motions of bodies with strong self-gravity are the same as those with weak self-gravity. In the case of J0337+1715, a neutron star with a dimensionless gravitational binding energy $3GM/5Rc^2 \approx 0.1$ and a low-mass white dwarf with much smaller gravitational binding energy ($\sim 3 \times 10^{-6}$) are both falling in the relatively strong gravitational potential of the outer white dwarf. The difference in the gravitational binding energies of the pulsar and the white dwarf of five orders of magnitude, as well as the large absolute value for the pulsar, provide a much larger 'lever arm' for testing the SEP than do Solar System tests, where the planets and moons have gravitational binding energies between 10^{-11} and 10^{-9}. Previous strong-field SEP tests used MSP/white dwarf systems and the Galactic field as the external perturbing field[28,29]. In the case of J0337+1715, the perturbing field (that of the outer white dwarf) is six to seven orders of magnitude larger, greatly magnifying any possible SEP violation effects[26,27]. Because most metric theories of gravity apart from general relativity predict violations of the SEP at some level, high-precision timing of J0337+1715 should soon produce unique and extremely interesting new tests of gravity[26].

METHODS SUMMARY

We have taken many hundreds of hours of radio timing observations with the GBT, the Arecibo telescope and the WSRT over the past 2 yr, with the best observations having time-of-arrival uncertainties of 0.8 μs in 10 s of data. These TOAs are fitted using a high-precision numerical integrator, including Newtonian gravitational effects from the three-body interactions as well as special relativistic transverse Doppler corrections and general relativistic Einstein and Shapiro delays (the last two are not yet important for the fitting). Uncertainties on the fitted parameters are derived using MCMC techniques.

The optical, ultraviolet and infrared photometry of the inner object was fitted using an absorbed white dwarf atmosphere, yielding results very close to the spectroscopic values. On the basis of spectroscopic gravity, the photometric radius of the inner white dwarf is $(0.091 \pm 0.005)R_{\odot}$ ($R_{\odot}$, solar radius), which leads to a photometric distance to the system. We see no emission from the outer object and can reject all single or binary main-sequence stars as being the outer companion. The data are consistent with a $0.4M_{\odot}$ white dwarf.

The radio timing fits benefitted from a radio interferometric position of J0337+1715 determined from a 3-h observation with the VLBA. The absolute positional accuracy is estimated to be 1–2 mas. A series of observations that has already begun will determine the parallax distance to a precision of 1%–2% as well

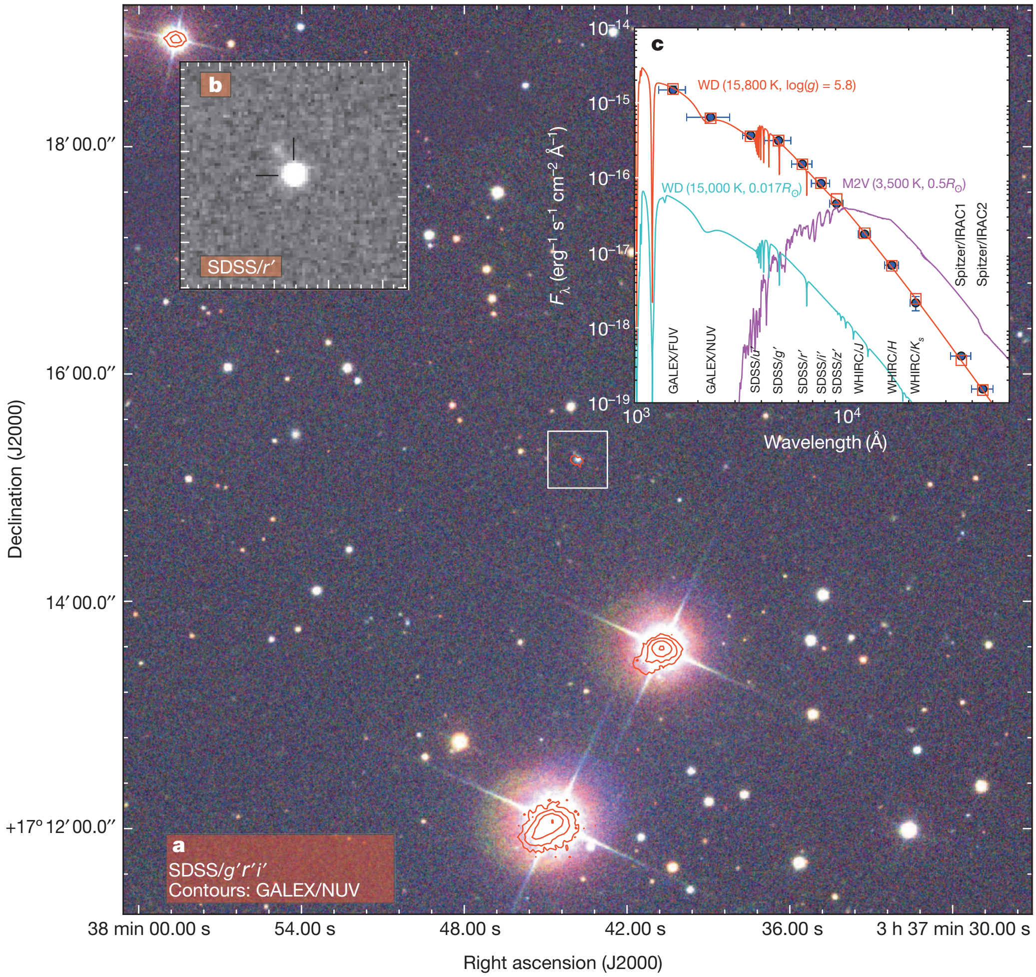


Figure 3 | Optical, infrared and ultraviolet data on PSR J0337+1715. **a**, Three-colour optical image around J0337+1715 from the SDSS (Methods). The contours are Galaxy Evolution Explorer (GALEX) near-ultraviolet (NUV) data. The box at the centre is the area shown in **b**. **b**, 30′′ × 30′′ region of the SDSS r' filter (SDSS/r') image along with the VLBA position, indicated by the tick marks. **c**, Spectral energy distribution. The data from GALEX, SDSS, the WIYN High Resolution Infrared Camera (WHIRC) and the Spitzer Infrared Array Camera (IRAC) are the blue circles, as labelled. The error bars represent 1σ uncertainties in the y direction (flux density) and the widths of the photometric filters in the x direction (wavelength). The red curve is a model atmosphere consistent with our spectroscopic determination for the inner white dwarf (WD) companion. The red boxes are the model atmosphere integrated through the appropriate filter passbands. For comparison, we also plot two possible models for the outer companion: a M2V star[30] (magenta) and a $0.4M_{\odot}$ white dwarf (cyan). All models have been reddened with an extinction of $A_V = 0.44$ mag. The photometry is consistent with the light from the hot inner white dwarf (with a surface gravity $\log(g) = 5.8$) and a smaller and more massive outer white dwarf, but not a low-mass main-sequence star.

as the $237/D_{kpc}$-μas astrometric reflex motion on the sky caused by the outer orbit (D_{kpc}, distance to the system in kiloparsecs).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 October; accepted 26 November 2013.
Published online 5 January 2014.

1. Newton, I. *Philosophiae Naturalis Principia Mathematica* (Streater, 1687).
2. Gutzwiller, M. C. Moon-Earth-Sun: the oldest three-body problem. *Rev. Mod. Phys.* **70,** 589–639 (1998).
3. Tokovinin, A., Thomas, S., Sterzik, M. & Udry, S. Tertiary companions to close spectroscopic binaries. *Astron. Astrophys.* **450,** 681–693 (2006).
4. Rappaport, S. *et al.* Triple-star candidates among the Kepler binaries. *Astrophys. J.* **768,** 33 (2013).
5. Fabrycky, D. C. in *Exoplanets* (ed. Seager, S.) 217–238 (Univ. Arizona Press, 2011).
6. Kopeikin, S. & Vlasov, I. Parametrized post-Newtonian theory of reference frames, multipolar expansions and equations of motion in the *N*-body problem. *Phys. Rep.* **400,** 209–318 (2004).
7. Thorsett, S. E., Arzoumanian, Z., Camilo, F. & Lyne, A. G. The triple pulsar system PSR B1620–26 in M4. *Astrophys. J.* **523,** 763–770 (1999).
8. Sigurdsson, S., Richer, H. B., Hansen, B. M., Stairs, I. H. & Thorsett, S. E. A young white dwarf companion to pulsar B1620–26: evidence for early planet formation. *Science* **301,** 193–196 (2003).
9. Bhattacharya, D. & van den Heuvel, E. P. J. Formation and evolution of binary and millisecond radio pulsars. *Phys. Rep.* **203,** 1–124 (1991).
10. Boyles, J. *et al.* The Green Bank Telescope 350 MHz drift-scan survey. I. Survey observations and the discovery of 13 pulsars. *Astrophys. J.* **763,** 80 (2013).
11. Lynch, R. S. *et al.* The Green Bank Telescope 350 MHz drift-scan survey II: data analysis and the timing of 10 new pulsars, including a relativistic binary. *Astrophys. J.* **763,** 81 (2013).
12. Wolszczan, A. & Frail, D. A. A planetary system around the millisecond pulsar PSR1257+12. *Nature* **355,** 145–147 (1992).
13. Wolszczan, A. confirmation of Earth-mass planets orbiting the millisecond pulsar PSR B1257+12. *Science* **264,** 538–542 (1994).
14. Backer, D. C. & Hellings, R. W. Pulsar timing and general relativity. *Annu. Rev. Astron. Astrophys.* **24,** 537–575 (1986).
15. Peale, S. J. On the verification of the planetary system around PSR 1257+12. *Astron. J.* **105,** 1562–1570 (1993).
16. Abazajian, K. N. *et al.* The seventh data release of the Sloan Digital Sky Survey. *Astrophys. J. Suppl. Ser.* **182,** 543–558 (2009).
17. Portegies Zwart, S., van den Heuvel, E. P. J., van Leeuwen, J. & Nelemans, G. The formation of the eccentric-orbit millisecond pulsar J1903+0327 and the origin of single millisecond pulsars. *Astrophys. J.* **734,** 55 (2011).
18. Champion, D. J. *et al.* An eccentric binary millisecond pulsar in the Galactic plane. *Science* **320,** 1309–1312 (2008).
19. Freire, P. C. C. *et al.* On the nature and evolution of the unique binary pulsar J1903+0327. *Mon. Not. R. Astron. Soc.* **412,** 2763–2780 (2011).
20. Mardling, R. A. New developments for modern celestial mechanics - I. General coplanar three-body systems. Application to exoplanets. *Mon. Not. R. Astron. Soc.* **435,** 2187–2226 (2013).

21. Ford, E. B., Kozinsky, B. & Rasio, F. A. Secular evolution of hierarchical triple star systems. *Astrophys. J.* **535,** 385–401 (2000).
22. Larwood, J. D. & Papaloizou, J. C. B. The hydrodynamical response of a tilted circumbinary disc: linear theory and non-linear numerical simulations. *Mon. Not. R. Astron. Soc.* **285,** 288–302 (1997).
23. Tauris, T. M. & Savonije, G. J. Formation of millisecond pulsars. I. Evolution of low-mass X-ray binaries with $P_{orb} > 2$ days. *Astron. Astrophys.* **350,** 928–944 (1999).
24. Özel, F., Psaltis, D., Narayan, R. & Santos Villarreal, A. On the mass distribution and birth masses of neutron stars. *Astrophys. J.* **757,** 55 (2012).
25. Lattimer, J. M. The nuclear equation of state and neutron star masses. *Annu. Rev. Nucl. Particle Sci.* **62,** 485–515 (2012).
26. Will, C. M. The confrontation between general relativity and experiment. *Living Rev. Relativ.* **9,** 3 (2006).
27. Freire, P. C. C., Kramer, M. & Wex, N. Tests of the universality of free fall for strongly self-gravitating bodies with radio pulsars. *Classical Quant. Grav.* **29,** 184007 (2012).
28. Stairs, I. H. *et al.* Discovery of three wide-orbit binary pulsars: implications for binary evolution and equivalence principles. *Astrophys. J.* **632,** 1060–1068 (2005).
29. Gonzalez, M. E. *et al.* High-precision timing of five millisecond pulsars: space velocities, binary evolution, and equivalence principles. *Astrophys. J.* **743,** 102 (2011).
30. Kurucz, R. *ATLAS9 Stellar Atmosphere Programs and 2 km/s Grid* (Kurucz CD-ROM No. 13., Smithsonian Astrophysical Observatory, 1993).

Acknowledgements We thank D. Levitan and R. Simcoe for providing optical and infrared observations; J. Deneva for early Arecibo telescope observations; P. Bergeron for use of his white dwarf photometry models; K. O'Neil and F. Camilo for approving discretionary time observations on the GBT and the Arecibo telescope, respectively; J. Heyl, E. Algol, and P. Freire for discussions; and G. Kuper, J. Sluman, Y. Tang, G. Jozsa, and R. Smits for their help supporting the WSRT observations. The GBT and VLBA are operated by the National Radio Astronomy Observatory, a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. The Arecibo Observatory is operated by SRI International in alliance with Ana G. Méndez-Universidad Metropolitana and the Universities Space Research Association, under a cooperative agreement with the National Science Foundation. The WSRT is operated by the Netherlands Institute for Radio Astronomy (ASTRON). This paper made use of data from the WIYN Observatory at Kitt Peak National Observatory, National Optical Astronomy Observatory, which is operated by the Association of Universities for Research in Astronomy under cooperative agreement with the National Science Foundation. This work is also based in part on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. I.H.S., V.M.K., M.H.v.K. and A.B. acknowledge support from NSERC. A.M.A. and J.W.T.H. acknowledge support from a Vrije Competitie grant from NWO. J.B., D.R.L, V.I.K. and M.A.M. were supported by a WV EPSCoR Research Challenge Grant. V.M.K. acknowledges support from CRAQ/FQRNT, CIFAR, the Canada Research Chairs Program and the Lorne Trottier Chair.

Author Contributions S.M.R., M.A.M. and D.R.L. were joint principal investigators of the GBT survey that found the pulsar, and all other authors except D.L.K., M.H.v.K., A.T.D., S.C. and A.S.-R. were members of the survey team who observed and processed data. J.B. found the pulsar in the search candidates. S.M.R. identified the source as a triple, wrote follow-up proposals, observed with the GBT, phase-connected the timing solution and wrote the manuscript. I.H.S. and J.W.T.H. performed timing observations, wrote follow-up proposals and substantially contributed to the initial timing solution. A.M.A. developed the successful timing model and performed the numerical integrations and MCMC analyses. D.L.K. identified the optical counterpart and then, with M.H.v.K. and A.S.-R., performed optical and infrared observations and the multiwavelength analysis. M.H.v.K. and D.L.K. both helped develop parts of the timing model. A.T.D. and S.C. performed the VLBA analysis. All authors contributed to interpretation of the data and the results and to the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.M.R. (sransom@nrao.edu).

METHODS

Radio timing observations. Continuing timing observations over most of the past two years have been undertaken with Arecibo every few weeks, the GBT every week to 10 days and the WSRT nearly daily, providing complementary timing precision and cadence. Without a complete timing solution, nor even a rough estimate of outer-orbit parameters for most of the duration, the rapid cadence was essential to maintain unambiguous count of the pulsar's rotations. Similar centre frequencies (1,440 MHz and 1,500 MHz), effective bandwidths ($\sim$600 MHz and $\sim$700 MHz) and observing systems[31] (PUPPI and GUPPI) are used at Arecibo and the GBT, respectively, whereas at the WSRT, the PUMAII instrument[32] is used with $\sim$160 MHz of bandwidth centred at 1,380 MHz. In all cases, the data are coherently de-dispersed[33] at the pulsar's dispersion measure (21.3162(3) pc cm^{-3}) and folded modulo estimates of the predicted apparent pulsar spin period, initially determined from a polynomial expansion of inner-orbital parameters refitted on a weekly to monthly basis.

Pulse TOAs, determined by cross-correlating high-signal-to-noise template profiles with folded pulse profiles using standard procedures[34], are measured every 10 s to 10 min depending on the telescope. We use the precise VLBA position (see below) and TEMPO2[35] to convert the arrival times at the observatory locations to the Solar System barycentre using the JPL DE405 planetary system ephemeris.

Timing fitting procedures and the three-body model. High-precision three-body integrations determine the stellar masses and a nearly complete orbital geometry using only well-tested Newtonian gravity and special relativity. For each set of trial parameters, we compute the pulsar and companion masses, positions and velocities, and then use Newtonian gravity to compute their accelerations. We evolve the system forward using a Bulirsch–Stoer differential equation solver[36] (using 80-bit floating-point precision with ODEINT in the Boost library), obtaining position accuracy (limited by round-off and truncation errors) of the order of 1 m. We then compute the Rømer and Einstein delays and use a spin-down model of the pulsar to produce a set of predicted TOAs, which we compare with the observed TOAs using a weighted sum of squared residuals.

Fits to the measured TOAs without obvious systematic residuals are impossible without the inclusion of the three-body interactions as well as the special relativistic transverse Doppler effect. General relativity is, in general, unimportant in the fitting of the system, but we calculate the full Einstein and Shapiro delays[14] based on the determined system masses and geometry and incorporate them into the resulting best-fit parameters. Ignoring these effects would lead to a distortion of orbital parameters, particularly the projected semimajor axes, because the delays would be absorbed into the fit. The magnitudes of the Shapiro delays peak-to-peak over the inner and outer orbits are $\sim$2.9 μs and $\sim$5.8 μs, respectively.

The parameter space is explored using MCMC techniques[37], and the parameter values given in Table 1 are the Bayesian posterior expected values. We also use the posterior distribution to compute the standard deviations, quoted as 1σ uncertainties. This process marginalizes over covariances between parameters and derived values.

Ultraviolet, optical, and infrared observations. After we identified J0337+1715 in SDSS data release 7[16], we identified the same object as an ultraviolet source in the GALEX All-sky Imaging Survey[38], confirming the blue colours (Fig. 3). We obtained further near-infrared photometry with the WHIRC imager[39] on the Wisconsin Indiana Yale NOAO (WIYN) 3.5-m telescope and mid-infrared photometry with the post-cryogenic Infrared Array Camera (IRAC) on board the Spitzer Space Telescope[40].

We fitted the data using synthetic white dwarf photometry[41] (extended to GALEX bands by P. Bergeron, personal communication), finding extinction $A_V = 0.34 \pm 0.04$ mag and effective temperature $T_{\rm eff} = 14{,}600 \pm 400$ K, with $\chi^2 = 7.7$ for nine degrees of freedom (Fig. 3c). This is close to the effective temperature we determined via optical spectroscopy (D.L.K. *et al.*, manuscript in preparation), $T_{\rm eff,spec} = 15{,}800 \pm 100$ K. Radial velocity measurements from those spectroscopic observations confirm that the optical star is the inner white dwarf in the system. Given the spectroscopy-determined temperature and surface gravity of $\log(g) = 5.82 \pm 0.05$, and a radius of $(0.091 \pm 0.005)R_\odot$ based on the white dwarf mass from pulsar timing ($0.197M_\odot$), the synthetic photometry provides a photometric distance to the system of $1{,}300 \pm 80$ pc. That distance is somewhat larger than the $\sim$750 pc implied by the observed dispersion measure towards the pulsar and the NE2001 Galactic free-electron-density model[42], although the latter probably has a large error range.

The photometry we measure is fully consistent with expectations for the inner companion only, as seen in Fig. 3. No additional emission is needed over the 1,000–50,000 Å range, although only where we actually have spectra can we be certain that no other emission is present. Given its known mass of $0.4M_\odot$, if the outer companion were a main-sequence star we would expect a spectral type of roughly M2V[43], implying an effective temperature near 3,500 K and a radius of $0.5R_\odot$. Figure 3 shows the results of such a stellar model[30], which exceed the near-infrared and mid-infrared data by a factor of >5, ruling out a main-sequence star as the outer companion. Two main-sequence $0.2M_\odot$ stars would also cause an excess in the near- and mid-infrared by a factor of $\sim$2. Instead we find that a $0.4M_\odot$ white dwarf with an effective temperature of $<$20,000 K (using synthetic photometry[41] again) is almost certainly the outer companion: such an object with $\log(g) = 7.5$ and radius $0.018R_\odot$ leads to a change in χ^2 of 1 compared with the fit for only a single photosphere. An example of such an outer companion is also shown in Fig. 3.

VLBA observations. The position of J0337+1715 used in the timing analysis was determined from a single 3-h observation with the VLBA on 13 February 2013, the first in a series of astrometric observations that will ultimately provide a $\sim$1%–2% parallax distance and transverse velocity for the system. Eight dual-polarization, 32-MHz-wide sub-bands were sampled from within the range 1,392–1,712 MHz, avoiding strong sources of radio-frequency interference. The bright source J0344+1559 was used as a primary phase reference source, and a phase referencing cycle time of 4.5 min (total cycle) was employed.

The multi-field correlation capability of the DiFX software correlator used at the VLBA[44] made it possible to inspect all catalogued sources from the NRAO VLA Sky Survey[45] that fell within the VLBA field of view; of the 44 such sources, four were detected by the VLBA and J033630.1+172316 was found to be a suitable secondary calibrator, with a peak flux density of 4 mJy per beam. The use of an 'in-beam' secondary calibrator reduces the spatial and temporal interpolation of the calibration solutions and improves the (relative) astrometric precision substantially[46]. J0337+1715 was detected with a signal-to-noise ratio of 30, providing a formal astrometric precision of around 0.1 mas.

The absolute positional accuracy of J0337+1715 in the International Celestial Reference Frame is currently limited by the registration of the position of J033630.1+172316 relative to J0344+1559; given the angular separation of 2.3° and the single observation, this is estimated to be 1–2 mas. This uncertainty in the absolute position will be reduced by additional VLBA observations. In addition to improving the absolute position and solving for parallax and proper motion, a full VLBA astrometric model will incorporate the $237/D_{\rm kpc}$-μas reflex motion on the sky caused by the outer orbit.

31. DuPlain, R. *et al.* in *Advanced Software and Control for Astronomy II* (eds Bridger, A. & Radziwill, N. M.) (SPIE Conf. Ser. 7019, SPIE, 2008).
32. Karuppusamy, R., Stappers, B. & van Straten, W. PuMa-II: a wide band pulsar machine for the Westerbork Synthesis Radio Telescope. *Proc. Astron. Soc. Pacif.* **120,** 191–202 (2008).
33. Hankins, T. H. & Rickett, B. J. in *Methods in Computational Physics* Vol. 14 *Radio Astronomy* (eds Alder, B., Fernbach, S. & Rotenberg, M.) 55–129 (Academic, 1975).
34. Taylor, J. H. Pulsar timing and relativistic gravity. *R. Soc. Lond. Phil. Trans. A* **341,** 117–134 (1992).
35. Hobbs, G. B., Edwards, R. T. & Manchester, R. N. TEMPO2, a new pulsar-timing package - I. An overview. *Mon. Not. R. Astron. Soc.* **369,** 655–672 (2006).
36. Bulirsch, R. & Stoer, J. Asymptotic upper and lower bounds for results of extrapolation methods. *Numer. Math.* **8,** 93–104 (1966).
37. Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. emcee: The MCMC Hammer. *Proc. Astron. Soc. Pacif.* **125,** 306–312 (2013).
38. Morrissey, P. *et al.* The calibration and data products of GALEX. *Astrophys. J. Suppl. Ser.* **173,** 682–697 (2007).
39. Meixner, M. *et al.* Design overview and performance of the WIYN High Resolution Infrared Camera (WHIRC). *Proc. Astron. Soc. Pacif.* **122,** 451–469 (2010).
40. Fazio, G. G. *et al.* The Infrared Array Camera (IRAC) for the Spitzer Space Telescope. *Astrophys. J. Suppl. Ser.* **154,** 10–17 (2004).
41. Tremblay, P.-E., Bergeron, P. & Gianninas, A. An improved spectroscopic analysis of DA white dwarfs from the Sloan Digital Sky Survey data release 4. *Astrophys. J.* **730,** 128 (2011).
42. Cordes, J. M. & Lazio, T. J. W. NE2001.I. A new model for the Galactic distribution of free electrons and its fluctuations. Preprint at http://arxiv.org/abs/astroph/0207156 (2002).
43. Cox, A. N. *Allen's Astrophysical Quantities* 4th edn, 381–396 (AIP Press/Springer, 2000).
44. Deller, A. T. *et al.* DiFX-2: a more flexible, efficient, robust, and powerful software correlator. *Proc. Astron. Soc. Pacif.* **123,** 275–287 (2011).
45. Condon, J. J. *et al.* The NRAO VLA Sky Survey. *Astron. J.* **115,** 1693–1716 (1998).
46. Chatterjee, S. *et al.* Precision astrometry with the Very Long Baseline Array: parallaxes and proper motions for 14 pulsars. *Astrophys. J.* **698,** 250–265 (2009).

doi:10.1038/nature12918

Localized sources of water vapour on the dwarf planet (1) Ceres

Michael Küppers[1], Laurence O'Rourke[1], Dominique Bockelée-Morvan[2], Vladimir Zakharov[2], Seungwon Lee[3], Paul von Allmen[3], Benoît Carry[1,4], David Teyssier[1], Anthony Marston[1], Thomas Müller[5], Jacques Crovisier[2], M. Antonietta Barucci[2] & Raphael Moreno[2]

The 'snowline' conventionally divides Solar System objects into dry bodies, ranging out to the main asteroid belt, and icy bodies beyond the belt. Models suggest that some of the icy bodies may have migrated into the asteroid belt[1]. Recent observations indicate the presence of water ice on the surface of some asteroids[2–4], with sublimation[5] a potential reason for the dust activity observed on others. Hydrated minerals have been found[6–8] on the surface of the largest object in the asteroid belt, the dwarf planet (1) Ceres, which is thought to be differentiated into a silicate core with an icy mantle[9–11]. The presence of water vapour around Ceres was suggested by a marginal detection of the photodissociation product of water, hydroxyl (ref. 12), but could not be confirmed by later, more sensitive observations[13]. Here we report the detection of water vapour around Ceres, with at least 10^{26} molecules being produced per second, originating from localized sources that seem to be linked to mid-latitude regions on the surface[14,15]. The water evaporation could be due to comet-like sublimation or to cryo-volcanism, in which volcanoes erupt volatiles such as water instead of molten rocks.

We observed Ceres with the Heterodyne Instrument for the Far Infrared (HIFI)[16] on the European Space Agency's Herschel Space Observatory[17] on four occasions between November 2011 and March 2013 (Extended Data Table 1) as part of the MACH-11 ('Measurements of 11 asteroids and comets with Herschel') guaranteed time programme (principal investigator L.O'R.) and of a follow-up Director's Discretionary Time Program. We used HIFI to search for water vapour directly, because it is more sensitive to water concentrated in the near-Ceres environment than previous instruments used to search for hydroxyl (OH). We observed the water ground-state line at a frequency of 556.936 GHz. The angular diameter of Ceres was <1 arcsec for all observations, compared to the beam width of HIFI, which was approximately 40 arcsec at the frequency of the water line. Although we cannot resolve Ceres spatially, we can derive information about the longitudinal distribution of the water sources on the surface from the variation of the absorption over the rotation of Ceres. Details of observations and data reduction are provided in the Supplementary Information and in Extended Data Table 1.

Figure 1 shows time-averaged spectra taken in October 2012 and on 6 March 2013, normalized to the thermal continuum of Ceres (measured with the expected brightness, see Extended Data Table 2). At the frequency of the water line, absorption in the thermal continuum of Ceres is clearly visible in the late 2012 observations, whereas in the 2013 data it is next to a weaker emission line detected at the 3σ level. The low outflow velocity (0.3–0.7 km s^{-1}) determined from the offset of the absorption line is comparable to the escape velocity of Ceres (about 0.52 km s^{-1}; ref. 18), showing that a fraction of the evaporated water does not escape from Ceres. For line strengths and offset information, see Extended Data Table 3.

The strength of the absorption is variable on short timescales (hours; Fig. 2) as well as on longer timescales (weeks and months; Extended Data Fig. 1 and Extended Data Table 3). We interpret the short-term variation in terms of localized sources on Ceres rotating into and out of the hemisphere visible by Herschel. Figure 2 shows the correlation of the strength of the absorption line with the position of features on the

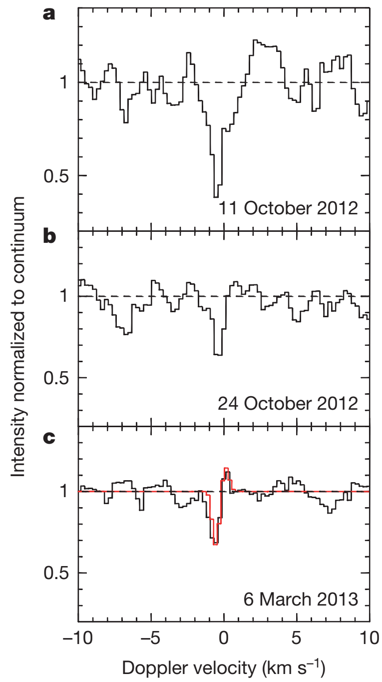


Figure 1 | Submillimetre water absorption line from the dwarf planet (1) Ceres. The spectra of the ground-state transition line 1_{10}–1_{01} of ortho-water at 556.939 GHz were obtained on 11.83–11.92 October 2012 UT (**a**), 24.84–24.96 October 2012 UT (**b**) and 6.13–6.55 March 2013 UT (**c**), with HIFI's Wide-Band Spectrometer. The spectra, which are the averages of the linear H and V polarizations, were divided by the Ceres continuum thermal emission. The abscissa represents the Doppler velocity in the Ceres frame, after correction for the relative motion between Ceres and Herschel. The spectral resolution is 1.1 MHz (0.5 km s^{-1}) with 0.6 MHz sampling. The water line is seen in absorption against the thermal emission of Ceres. Material moving towards the observer causes the absorption line to be blue-shifted. In the 6 March spectrum (**c**), a redshifted emission line is visible next to the blue-shifted absorption line, showing that the exosphere of Ceres extends towards the limbs. The possible polarization of this line is discussed in the Supplementary Information. Overplotted on the 6 March spectrum is a model of the spectrum of the water line for two active spots 60 km in diameter situated on the surface of Ceres (red spectrum in **c**). The simulation takes into account the variation of the sub-observer point longitude during the 10-hour-long observation. The model spectrum is adjusted to the depth of the observed spectrum. The relative strengths of the redshifted and blue-shifted peaks are correctly reproduced.

[1]European Space Agency, European Space Astronomy Centre, PO Box 78, Villanueva de la Cañada 28691, Spain. [2]Laboratoire d'études spatiales et d'instrumentation en astrophysique, Observatoire de Paris, CNRS, Université Pierre et Marie Curie (UPMC), Université Paris-Diderot, 5 Place Jules Janssen, 92195 Meudon, France. [3]Jet Propulsion Laboratory, Pasadena, 4800 Oak Grove Drive, La Cañada Flintridge, California 91011, USA. [4]Institut de Mécanique Céleste et de Calcul des Éphémérides, Observatoire de Paris, Unité Mixte de Recherche (UMR) 8028. CNRS, 77 Avenue Denfert Rochereau, 75014 Paris, France. [5]Max-Planck-Institut für extraterrestrische Physik (MPE), Giessenbachstrasse 1, 85748 Garching, Germany.

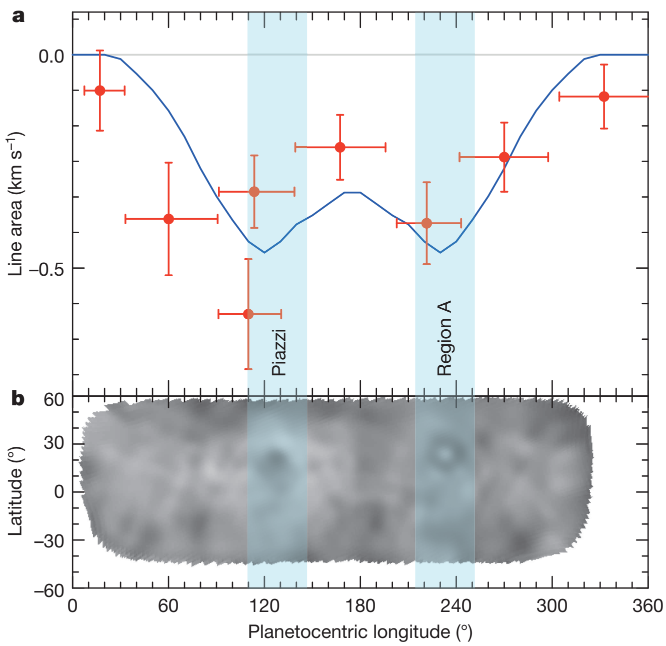


Figure 2 | Variability of water absorption on 6 March 2013. **a**, Line area of the water absorption line (normalized to the continuum emission of Ceres) at 557 GHz as a function of the longitude of the sub-observer point. Measurements are shown as red dots; error bars on the intensity are 1σ and the horizontal bars show the range of sub-observer longitudes covered by individual measurements. The two conflicting data points at sub-observer point longitude $\Lambda \approx 110^\circ$ were taken within a time interval of 9 hours (corresponding to the rotation period of Ceres), and suggest temporal variability at the regional scale. Vertical bands indicate the planetocentric longitude of the dark regions: Piazzi (longitude, 123°, latitude +21°) and Region A (longitude 231°, latitude +23°)[14,15,19]. The curve in blue is the result of a gas-kinetic model of the exosphere of Ceres[21] (see Supplementary Information). Water is released from localized sources 60 km in diameter situated at the longitudes and latitudes of regions Piazzi and Region A, with a total production rate of 10^{26} molecules per second for each source. The surface temperature of Ceres varies from 235 K (subsolar, that is, when the Sun is at zenith) to 168 K (morning and evening). The excitation and radiative transfer models of the water 1_{10}–1_{01} line include excitation of the vibrational bands by the Sun's infrared radiation, excitation of the rotational lines by thermal radiation from Ceres, collisions with water and self-absorption effects[22] (see Supplementary Information). **b**, A map of Ceres from near-infrared adaptive-optics imaging observations[14]. Piazzi and Region A are seen as dark regions, with a bright centre within Region A.

Ceres surface that are known from ground-based[14,15] and Hubble Space Telescope[19] observations. In all observations that detected water vapour from Ceres, the absorption line strength is strongly correlated with the visibility of surface areas identified as dark regions (about 5% darker than the average surface) in near-infrared observations. We identify those regions as the likely source of most of the evaporating water. A bright region known from observations in the visible region of the spectrum does not appear to contribute. Possibly, the dark regions are warmer than the average surface, resulting in efficient sublimation of small water-ice reservoirs.

Although the small number of observations does not allow a unique interpretation of the long-term variation, the lack of detection of the water line at 2.94 astronomical units (AU; where 1 AU is the mean distance from Earth to the Sun) in November 2011 and its first detection at 2.72 AU are consistent with the steep increase of water-ice sublimation between 3 AU and 2.5 AU (ref. 20). In addition, the larger absorption strength on 11 October 2012 compared to the observations two weeks later and five months later suggests sporadic changes in the water evaporation. Given that the spin axis of Ceres is nearly perpendicular to its orbital plane[14], we expect seasonal variations driven by spin-axis obliquity to contribute little to the variability.

We analysed the water exosphere of Ceres with a gas kinetic Direct Simulation Monte Carlo[21] model (Extended Data Fig. 2) that considers water vapour to be ejected from localized sources, and then to slow down in Ceres' gravity field. To simulate water spectra, we use a state-of-the-art two-dimensional excitation model[22], which considers excitation by radiation from Ceres and the Sun and collisional excitation (see details in Supplementary Information). The temporal variation of the absorption line observed on 6 March 2013 is well described by a model that considers outgassing from two sources coincident with dark regions Piazzi and Region A (Fig. 2). Modelling predicts line emission at positive velocities (Fig. 1), caused by gas expansion from dense to more rarefied regions. The resulting total production rate of about 2×10^{26} molecules (or 6 kg) per second of water requires only a tiny fraction of the Ceres surface to be covered by water ice. The surface of Ceres receives on average a solar input power of approximately 50 W m^{-2} (a quarter of the total solar power at the heliocentric distance of Ceres, with the factor 1/4 being the ratio between the cross-section of Ceres and its surface area). Because Ceres is located in the transition range between the outer Solar System, where most of the solar energy will be re-emitted as thermal radiation, and the inner Solar System, where most of the energy will go into sublimation of the ice, we assume that half of the energy will be used for sublimation. With a latent heat of sublimation of 2.5×10^6 J kg^{-1}, the corresponding sublimation rate is 10^{-5} kg m^{-2} s^{-1}. To sublimate 6 kg s^{-1} of water ice, Ceres must have a surface area covered with water ice of 0.6 km^2, or approximately 10^{-7} of its total surface area. If the activity is restricted to areas with a radius of about 100 km (the approximate size of the identified source regions), the active surface fraction required within those areas is still very small ($<10^{-5}$ of the surface area of the identified source regions).

An unexpected aspect of the data is that the absorption line appears to be strongly linearly polarized in October 2012, whereas no significant polarization was seen in March 2013. See Extended Data Table 3, Extended Data Fig. 3, and Supplementary Information for further analysis.

The measured water production is two orders of magnitudes higher than is predicted from a model of sublimation maintained from water supplied from the interior of Ceres[23]. In addition, the water activity is most probably not concentrated on polar regions, where water ice would be most stable. We propose two mechanisms for maintaining the observed water production on Ceres. The first is cometary-type sublimation of (near) surface ice. In this case the sublimating ice drags near-surface dust with it and in this way locally removes the surface layer and exposes fresh ice. Transport from the interior is not required. The second mechanism is geysers or cryovolcanoes, for which an interior heat source is needed. For Jupiter's satellite Io and Saturn's moon Enceladus the source of activity is dissipation of tidal forces from the planet[24,25]. That can be excluded for Ceres, but some models suggest that a warm layer in the interior heated by long-lived radioisotopes may maintain cryovolcanism on Ceres at the present time (ref. 26 and references therein).

One way of distinguishing between the two mechanisms is to analyse the variation of the water activity of Ceres over its orbit. Taking the activity of main-belt comets as a reference, cometary activity is expected to be concentrated at the perihelion passage[5]. On the other hand, cryovolcanism receives its energy from the interior and so no dependence on heliocentric distance would be seen, although sporadic variations of activity are likely. The currently available data appear to be consistent with the cometary hypothesis, but more observations are needed to distinguish between these possibilities (see Fig. 3).

Although ground- and space-based observations may further map the behaviour of Ceres over its orbit, the Dawn spacecraft mission[27] arriving to orbit Ceres in early 2015 is expected to be key in providing a long-term follow-up on the water outgassing behaviour of Ceres. In particular, it will provide long-term monitoring of the water outgassing concentration and stability of the activity in the dark regions where we suggest that the water-ice mantle of Ceres may reach the surface. Two of the instruments on Dawn—the near-infrared spectrometer (VIR)

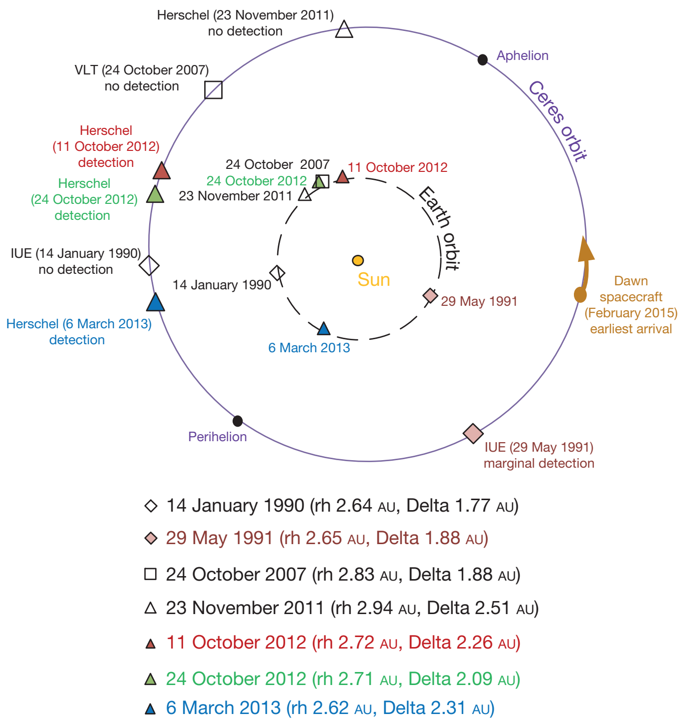


Figure 3 | Water production of Ceres versus position on its orbit. Searches for water activity on Ceres were performed with the International Ultraviolet Explorer (IUE), the Very Large Telescope (VLT), and Herschel. The inner orbit is that of Earth, the outer orbit that of Ceres. rh is the heliocentric distance of Ceres and Delta is the distance between Ceres and the observer. If cometary activity is the source of water on Ceres we would expect the onset of activity to appear well before perihelion before becoming much weaker at some time after perihelion. The pre-perihelion data are consistent with that picture. No activity was detected by VLT and Herschel at less than 2.83 AU; then Herschel detected activity in all observations within 2.72 AU. The non-detection by IUE at almost the same orbital position as one of the Herschel observations three orbital periods earlier can be explained by the higher sensitivity of Herschel for near-equatorial sources. The single observation postperihelion (a marginal detection by IUE) does not allow us to draw conclusions about the behaviour when Ceres is receding from the Sun. Dawn will visit Ceres on the postperihelion arc. The water absorption was strongest in the first Herschel detection on 11 October 2012, well before passing perihelion. To first order this is not what we would expect for cometary activity. It may have been caused by an analogue of a cometary outburst. Alternatively, it could have been a volcanic eruption. In that case, the correlation of the detectability with heliocentric distance may be coincidental. Additional observations are required to distinguish better between different mechanisms for the water activity.

and the gamma ray and neutron detector (GRaND)—may contribute significantly to this task. Although no observations of water are available for the orbital position of Ceres at the time of its arrival (Fig. 3) and the heliocentric distances in the spacecraft's initial few months around Dawn of 2.85–2.95 AU appear to be unfavourable for detecting activity, it may be that the post-perihelion activity is maintained to larger distances.

The identification of more than one water source on Ceres suggests outgassing from a small ice fraction near the surface as opposed to sporadic activity triggered by a singular event like a recent large impact. This supports the idea that Ceres possesses an icy mantle, and it also implies that we have detected water activity in the asteroid main belt. If the water is from cometary sublimation, it demonstrates that activity driven by water sublimation is not limited to classical comets, but is present in the asteroid belt as well. This supports the new vision of our Solar System with a continuum in composition and ice content between asteroid and comet populations[28].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 August; accepted 26 November 2013.

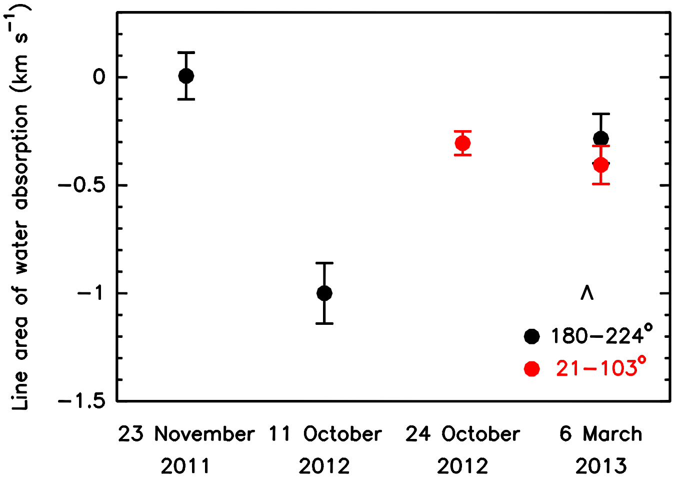
1. Walsh, K. J., Morbidelli, A., Raymond, S. N., O'Brien, D. P. & Mandell, A. M. A low mass for Mars from Jupiter's early gas-driven migration. *Nature* **475,** 206–209 (2011).
2. Campins, H. *et al.* Water ice and organics on the surface of the asteroid 24 Themis. *Nature* **464,** 1320–1321 (2010).
3. Rivkin, A. S. & Emery, J. P. Detection of ice and organics on an asteroidal surface. *Nature* **464,** 1322–1323 (2010).
4. Licandro, J. et al. (65) Cybele: detection of small silicate grains, water-ice, and organics. *Astron. Astrophys.* **525,** A34 (2011).
5. Jewitt, D. The active asteroids. *Astron. J.* **143,** 66 (2012).
6. Lebofsky, L. A., Feierberg, M. A., Tokunaga, A. T., Larson, H. P. & Johnson, J. R. The 1.7- to 4.2-micron spectrum of asteroid 1 Ceres: evidence for structural water in clay minerals. *Icarus* **48,** 453–459 (1981).
7. King, T. V. V., Clark, R. N., Calvin, W. M., Sherman, D. M. & Brown, R. H. Evidence for ammonium-bearing minerals on Ceres. *Science* **255,** 1551–1553 (1992).
8. Milliken, R. E. & Rivkin, A. S. Brucite and carbonate assemblages from altered olivine-rich materials on Ceres. *Nature Geosci.* **2,** 258–261 (2009).
9. Thomas, P. C. *et al.* Differentiation of the asteroid Ceres as revealed by its shape. *Nature* **437,** 224–226 (2005).
10. McCord, T. B. & Sotin, C. Ceres: evolution and current state. *J. Geophys. Res.* **110,** E05009 (2005).
11. Castillo-Rogez, J. C. & McCord, T. B. Ceres' evolution and present state constrained by shape data. *Icarus* **205,** 443–459 (2010).
12. A'Hearn, M. F. & Feldman, P. D. Water vaporization on Ceres. *Icarus* **98,** 54–60 (1992).
13. Rousselot, P. *et al.* A search for water vaporization on Ceres. *Astron. J.* **142,** 125 (2011).
14. Carry, B. *et al.* Near-infrared mapping and physical properties of the dwarf-planet Ceres. *Astron. Astrophys.* **478,** 235–244 (2008).
15. Carry, B. *et al.* The remarkable surface homogeneity of the Dawn mission target (1) Ceres. *Icarus* **217,** 20–26 (2012).
16. de Graauw, Th *et al.* The Herschel-Heterodyne Instrument for the Far-Infrared (HIFI). *Astron. Astrophys.* **518,** L6 (2010).
17. Pilbratt, G. L. *et al.* Herschel Space Observatory. An ESA facility for far-infrared and submillimetre astronomy. *Astron. Astrophys.* **518,** L1 (2010).
18. Carry, B. Density of asteroids. *Planet. Space Sci.* **73,** 98–118 (2012).
19. Li, J.-Y. *et al.* Photometric analysis of 1 Ceres and surface mapping from HST observations. *Icarus* **182,** 143–160 (2006).
20. Biver, N. *et al.* The 1995–2002 long-term monitoring of comet C/1995 O1 (Hale–Bopp) at radio wavelength. *Earth Moon Planets* **90,** 5–14 (2002).
21. Crifo, J. F., Loukianov, G. A., Rodionov, A. V. & Zakharov, V. V. Comparison between Navier–Stokes and direct Monte-Carlo simulations of the circumnuclear coma I. Homogeneous, spherical sources. *Icarus* **156,** 249–268 (2002).
22. Zakharov, V., Bockelée-Morvan, D., Biver, N., Crovisier, J. & Lecacheux, A. Radiative transfer simulation of water rotational excitation in comets. Comparison of the Monte Carlo and escape probability methods. *Astron. Astrophys.* **473,** 303–310 (2007).
23. Fanale, F. P. & Salvail, J. R. The water regime of asteroid (1) Ceres. *Icarus* **82,** 97–110 (1989).
24. Peale, S. J., Cassen, P. & Reynolds, R. T. Melting of Io by tidal dissipation. *Science* **203,** 892–894 (1979).
25. Howett, C. J. A., Spencer, J. R., Pearl, J. & Segura, M. High heat flow from Enceladus' south polar region measured using 10–600 cm-1 Cassini/CIRS data. *J. Geophys. Res.* **116,** E03003 (2011).
26. McCord, T. B., Castillo-Rogez, J. & Rivkin, A. Ceres: its origin, evolution and structure and *Dawn's* potential contribution. *Space Sci. Rev.* **163,** 63–76 (2011).
27. Russell, C. T. & Raymond, C. A. The Dawn mission to Vesta and Ceres. *Space Sci. Rev.* **163,** 3–23 (2011).
28. Gounelle, M. *et al.* in *The Solar System Beyond Neptune* (eds Barucci, M. A., Boehnhardt, H., Cruikshank, D. P. & Morbidelli, A.) 525–541 (Univ. Arizona Press, 2008).

Supplementary Information is available in the online version of the paper.

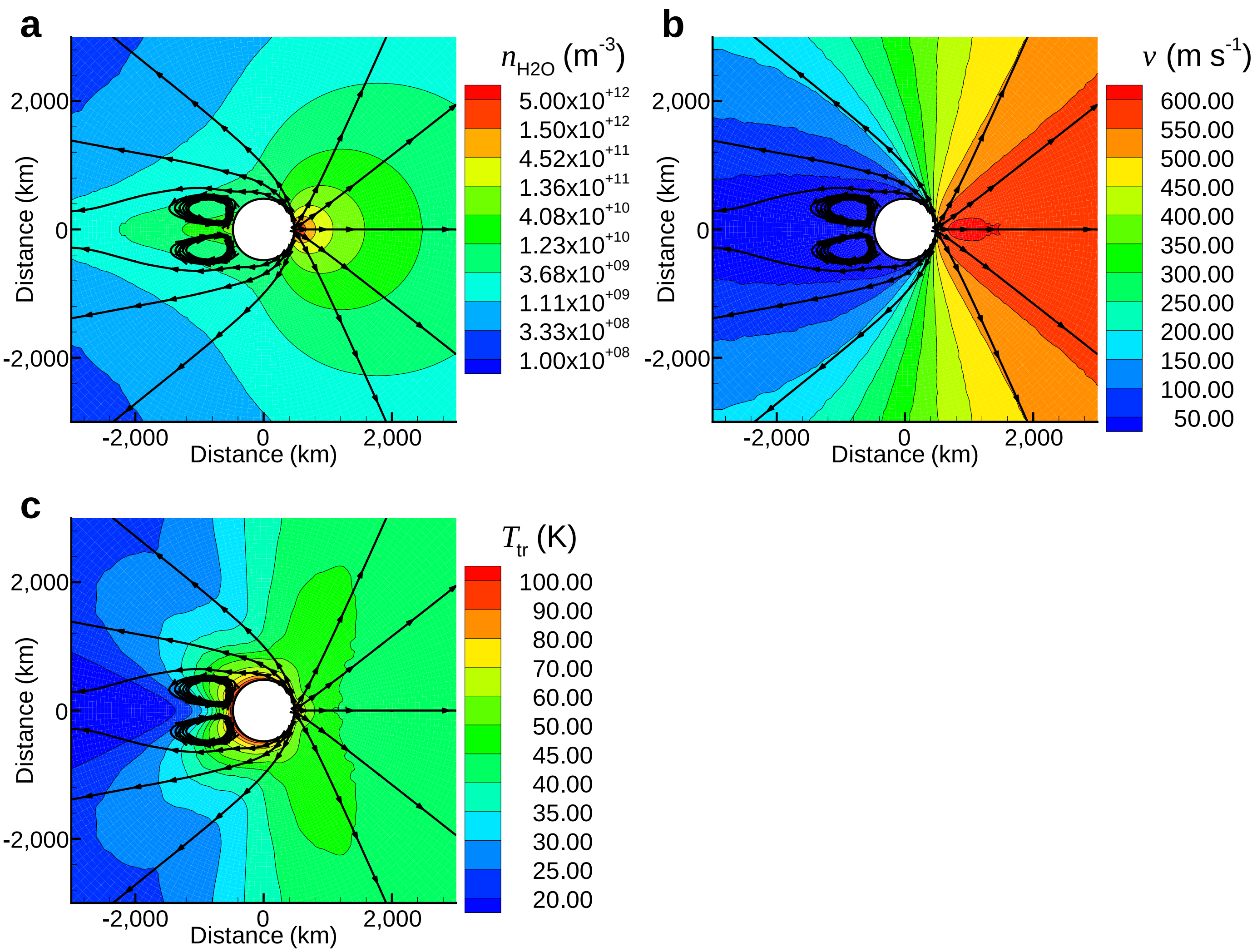
Acknowledgements Herschel is an ESA space observatory with science instruments provided by European-led principal investigator consortia and with important participation by NASA. The HIFI was designed and built by a consortium of institutes and university departments from across Europe, Canada and the United States under the leadership of SRON, the Netherlands Institute for Space Research, and with major contributions from Germany, France and the USA. This development was supported by national funding agencies: CEA, CNES, CNRS (France); ASI (Italy); and DLR (Germany). Additional funding support for some instrument activities was provided by the ESA. We thank the team at the Herschel Science Centre for their flexibility in scheduling the observations. We thank the Herschel Project Scientist and the Time Allocation Committee for the allocation of Director Discretionary Time. B.C. acknowledges support from the faculty of the European Space Astronomy Centre (ESAC). We thank A. Pollock for proofreading the final text.

Author Contributions M.K. proposed the observations of Ceres with HIFI as part of L.O'R.'s MACH-11 Guaranteed Time Program. M.K., L.O'R., D.B.-M., B.C., D.T. and A.M. planned the observations. M.K., D.B.-M., B.C., D.T., R.M. and J.C. contributed to the data analysis. The modelling was performed by D.B.-M., V.Z., S.L., P.v.A. and T.M. The manuscript was written by M.K., L.O'R., D.B.-M., B.C. and M.A.B. All authors discussed the results and reviewed the manuscript.

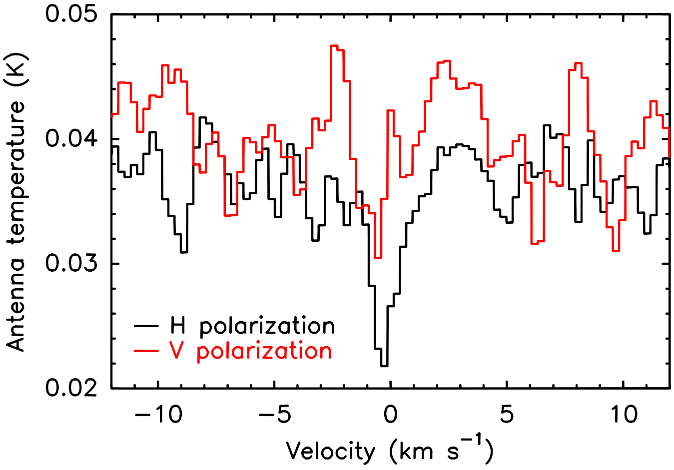
Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.K. (michael.kueppers@sciops.esa.int).



Extended Data Figure 1 | Long-term variability of water absorption. The absolute value of the area of the water absorption line at 557 GHz (line area normalized to Ceres continuum emission) is plotted for dates of observations covering the same sub-observer point longitudes: $\Lambda = 180^\circ$–204° on 23 November 2011, 11 October 2012 and 6 March 2013 (black dots); $\Lambda = 21^\circ$–103° on 24 October 2012 and 6 March 2013 (red dots). Error bars are 1σ. The strength of the absorption is variable on timescales of hours or months.



Extended Data Figure 2 | Direct Simulation Monte Carlo calculations of the exosphere of Ceres. **a–c,** Number density n_{H2O} (**a**), velocity v (**b**) and translational temperature T_{tr} (**c**) for water outgassing from an active spot about 60 km in diameter situated on the surface of Ceres at the subsolar point. The Sun is towards the right. The total water production rate is 10^{26} molecules per second. The Ceres surface temperature varies from 235 K (subsolar) to 168 K. See Supplementary Information. Stream lines are shown in black. The vortex seen on the night side is caused by the competition of molecules falling back on the surface owing to gravity and those molecules diffusing outwards. The local maximum in velocity observed above the active spot is also an effect of gravity. The gravity of Ceres causes 3% of the evaporated molecules to fall back to the surface, whereas 7% fall back owing to collisions between water molecules in the atmosphere.



Extended Data Figure 3 | The spectrum from 11 October 2012 in H and V polarizations. Although there is no significant polarization in the continuum, the line area is about 2.5 times larger in horizontal polarization than in the marginal detection of the line in vertical polarization.

Extended Data Table 1 | Overview of the acquired data

Start date and time (UT)	Duration (s)	r (AU)	Δ (AU)	Sub-observer point Λ (°)	ε (°)	Sub-solar point Λ_S (°)	ε_S (°)	Phase ϕ (°)
2011-11-23 11:31:20	4010	2.94	2.51	224-180	5	243-200	3	19
2012-10-11 19:53:46	8245	2.72	2.26	247-156	1	227-139	−1	21
2012-10-24 20:04:28	8575	2.71	2.09	103-8	1	84-350	−1	19
2013-03-06 03:05:42	8575	2.62	2.31	130-35	−7	152-57	−3	22
2013-03-06 05:30:03	9146	2.62	2.31	34-293	−7	56-318	−3	22
2013-03-06 08:03:55	9146	2.62	2.31	292-192	−7	317-214	−3	22
2013-03-06 10:37:47	9146	2.62	2.31	191-90	−7	213-111	−3	22

Geometric parameters of the observations are the heliocentric distance of Ceres r, the Ceres–Herschel distance Δ, the sub-observer point longitude Λ and subsolar point longitude Λ_S at the beginning and end of each observation[14,29], the sub-observer point latitude ε and subsolar point latitude ε_S (refs 14 and 29), and the phase angle ϕ. The Herschel observation identification numbers (Obsids) are 1342232694 (23 November 2011), 1342253122 (11 October 2012), 1342254428 (24 October 2012) and 1342266018–1342266021 (6 March 2013).

29. Chamberlain, M. A., Sykes, M. V. & Esquerdo, G. A. Ceres light curve analysis—period determination. *Icarus* **188,** 451–456 (2007).

Extended Data Table 2 | Continuum brightness in the spectra

Date (UT)		Measured continuum (Jy)	Expected continuum (Jy)
23.48-23.53	November 2011	7.24 ± 0.65	7.35 ± 0.4
11.83-11.92	October 2012	8.71 ± 1.16	9.78 ± 0.5
24.84-24.96	October 2012	11.48 ± 0.72	11.54 ± 0.6
6.13-6.55	March 2013	8.61 ± 0.62	8.74 ± 0.4

Measured and expected brightness of the continuum. The expected thermal continuum was calculated with a thermophysical model[30]. The estimated accuracy of the model is 5%.

30. Müller, T. G. *et al.* Herschel celestial calibration sources: four large main-belt asteroids as prime flux calibrators for the far-IR/sub-mm range. *Exp. Astron.* http://dx.doi.org/10.1007/s10686-013-9357-y (in the press).

Extended Data Table 3 | Characteristics of H_2O spectra

Date (UT)	Polarization	Line area ($km\ s^{-1}$)	Offset ($km\ s^{-1}$)	Width ($km\ s^{-1}$)	Absorbance (%)
23.48-23.53 November 2011	H+V	0.07± 0.10	-	-	< 10
11.83-11.92 October 2012	H	−1.07 ± 0.12	−0.26 ± 0.07	1.37 ± 0.21	74
	V	−0.43 ± 0.12	−0.76 ± 0.15	0.96 ± 0.26	42
	H+V	−0.79 ± 0.10	−0.40 ± 0.09	1.43 ± 0.22	52
24.84-24.96 October 2012	H	−0.56 ± 0.09	−0.72 ± 0.11	1.21 ± 0.19	44
	V	−0.22 ± 0.06	−0.16 ± 0.09	0.61 ± 0.18	34
	H+V	−0.31 ± 0.06	−0.39 ± 0.07	0.73 ± 0.16	39
6.13-6.55 March 2013	H	−0.27 ± 0.04	−0.67 ± 0.05	0.69 ± 0.10	37
	V	−0.20 ± 0.04	−0.69 ± 0.09	0.67 ± 0.18	28
	H+V	−0.24 ± 0.03	−0.68 ± 0.05	0.68 ± 0.09	33

Results of Gaussian fits to the water 1_{10}–1_{01} absorption line, after normalizing by the thermal continuum emission from Ceres. Offset refers to the radial velocity of the line centre relative to that of Ceres. The last column provides the absorbance (the percentage of the continuum that is absorbed) at the line centre from a Gaussian fit to the absorption line (1σ upper limit for the first observation). The V/H line area ratios are 0.40 ± 0.12, 0.38 ± 0.13 and 0.74 ± 0.20, for 11 October, 24 October and 6 March, respectively.

doi:10.1038/nature12800

Tunable symmetry breaking and helical edge transport in a graphene quantum spin Hall state

A. F. Young[1]*, J. D. Sanchez-Yamagishi[1]*, B. Hunt[1]*, S. H. Choi[1], K. Watanabe[2], T. Taniguchi[2], R. C. Ashoori[1] & P. Jarillo-Herrero[1]

Low-dimensional electronic systems have traditionally been obtained by electrostatically confining electrons, either in heterostructures or in intrinsically nanoscale materials such as single molecules, nanowires and graphene. Recently, a new method has emerged with the recognition that symmetry-protected topological (SPT) phases[1,2], which occur in systems with an energy gap to quasiparticle excitations (such as insulators or superconductors), can host robust surface states that remain gapless as long as the relevant global symmetry remains unbroken. The nature of the charge carriers in SPT surface states is intimately tied to the symmetry of the bulk, resulting in one- and two-dimensional electronic systems with novel properties. For example, time reversal symmetry endows the massless charge carriers on the surface of a three-dimensional topological insulator with helicity, fixing the orientation of their spin relative to their momentum[3,4]. Weakly breaking this symmetry generates a gap on the surface[5], resulting in charge carriers with finite effective mass and exotic spin textures[6]. Analogous manipulations have yet to be demonstrated in two-dimensional topological insulators, where the primary example of a SPT phase is the quantum spin Hall state[7,8]. Here we demonstrate experimentally that charge-neutral monolayer graphene has a quantum spin Hall state[9,10] when it is subjected to a very large magnetic field angled with respect to the graphene plane. In contrast to time-reversal-symmetric systems[7], this state is protected by a symmetry of planar spin rotations that emerges as electron spins in a half-filled Landau level are polarized by the large magnetic field. The properties of the resulting helical edge states can be modulated by balancing the applied field against an intrinsic antiferromagnetic instability[11–13], which tends to spontaneously break the spin-rotation symmetry. In the resulting canted antiferromagnetic state, we observe transport signatures of gapped edge states, which constitute a new kind of one-dimensional electronic system with a tunable bandgap and an associated spin texture[14].

In the integer quantum Hall effect, the topology of the bulk Landau-level energy bands requires the existence of gapless edge states at any interface with the vacuum. The metrological precision of the Hall quantization can be traced to the inability of these edge states to backscatter due to the physical separation of modes with opposite momenta by the insulating sample bulk. In contrast, counterpropagating boundary states in a SPT insulator coexist spatially but are prevented from backscattering by a symmetry of the experimental system[3,4]. The local symmetry that protects transport in SPT surface states is unlikely to be as robust as the inherently nonlocal physical separation that protects the quantum Hall effect. However, it enables the creation of new electronic systems in which momentum and some quantum number such as spin are coupled, potentially leading to devices with new functionality. Most experimentally realized SPT phases are based on time reversal symmetry (TRS), with counterpropagating states protected from intermixing by the Kramers degeneracy. However, intensive efforts are under way to search for topological phases protected by symmetries other than time reversal in new experimental systems[15,16].

Our approach is inspired by the similarity between the TRS quantum spin Hall (QSH) state and overlapping electron- and hole-like copies of the quantum Hall effect, with the two copies having opposite spin polarizations. This state is protected by spin conservation rather than the orthogonality of states in a Kramers doublet[3], as with the TRS QSH observed in systems with strong spin–orbit coupling. Nevertheless, it is expected to reproduce the characteristic experimental signatures of the TRS QSH, with gapless helical edge states enclosing an insulating bulk[9,10]. Two requirements are necessary for such a QSH state. First, the spin–orbit coupling must be weak, so that spin remains a good quantum number. Second, the energy gap between electron- and hole-like Landau levels must be small enough to be invertible by the Zeeman splitting. Both of these conditions are met in graphene, which is a gapless semimetal with very weak spin–orbit coupling[17]. The graphene Landau-level structure is characterized by the existence of a fourfold spin- and valley-degenerate Landau level at zero energy[18] (zLL). Near the sample boundary, the zLL splits into one positively dispersing (electron-like) and one negatively dispersing (hole-like) mode per spin projection. Consequently, a spin-symmetry-protected QSH state is expected when the spin degeneracy is lifted by an external magnetic field, resulting in a bulk energy gap at charge neutrality and electron- and hole-like states with opposite spin polarizations that cross at the sample edge[9,10].

Experimentally, charge-neutral monolayer graphene does not exhibit the expected phenomenology of the QSH state, becoming strongly insulating instead at high magnetic fields[19]. Although the precise nature of this insulating state has remained elusive, its origin can be traced to the strong Coulomb interactions within the graphene zLL. At integer filling factors, ν, the Coulomb energy is minimized by forming antisymmetric orbital wavefunctions, forcing the combined spin–valley isospin part of the wavefunction to be symmetric. The resulting possible ground states lie on a degenerate manifold of states fully polarized in the approximately SU(4)-symmetric isospin space[20], encompassing a variety of different spin and valley orders. In the real experimental system, the state at any given filling factor (such as $\nu = 0$) is determined by the competition between SU(4) symmetry-breaking effects. The most obvious such anisotropy is the Zeeman effect, which naturally favours a spin-polarized state, but the sublattice structure of the zLL adds additional interaction anisotropies[21] that can favour spin-unpolarized ground states characterized by lattice-scale spin- or charge-density-wave order[11–13,22]. This interplay can be probed experimentally by changing the in-plane component of the magnetic field, which changes the Zeeman energy but does not affect orbital energies, and previous observations indeed confirm that the state responsible for the $\nu = 0$ insulator is spin unpolarized[23]. However, the spin-polarized QSH state can be expected to emerge for a sufficiently large in-plane field, manifesting as an incompressible conducting state at charge neutrality.

Figure 1a shows two-terminal conductance measurements of a high-quality graphene device fabricated on a thin hexagonal boron nitride substrate, which itself sits on a graphite local gate. As the total magnetic field (B_T) is increased with $B_\perp$, the component perpendicular to the

[1]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Advanced Materials Laboratory, National Institute for Materials Science, 1-1 Namiki, Tsukuba 305-0044, Japan.
*These authors contributed equally to this work.

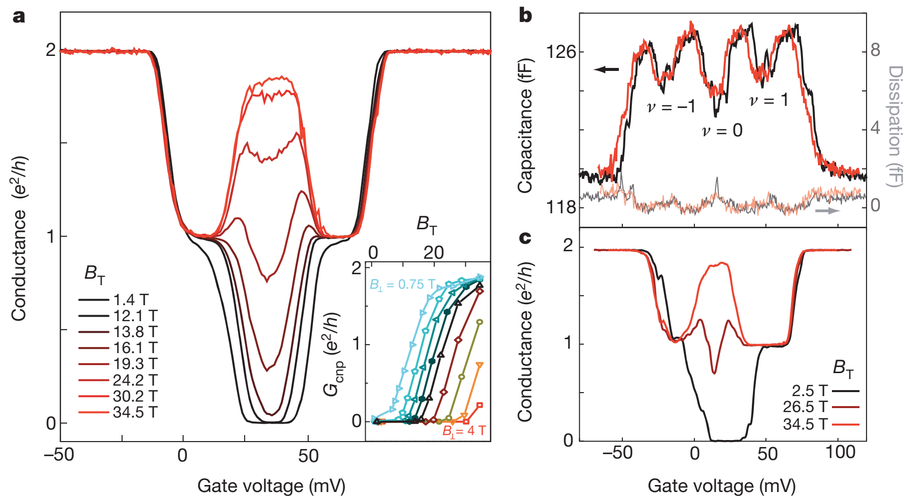


Figure 1 | QSH state in monolayer graphene in extreme tilted magnetic fields. **a**, Conductance of device A at $B_\perp = 1.4$ T for different values of B_T. As B_T increases, the insulating state at $\nu = 0$ is gradually replaced by a high-conductance state, with an accompanying inversion of the sign of $\partial G_{cnp}/\partial T$ (additional data in Extended Data Figs 2 and 3). Inset, G_{cnp} as a function of B_T for device A: $B_\perp = 0.75$, 1.0, 1.4, 1.6, 2.0, 2.5, 3.0 and 4.0 T (left to right). **b**, Capacitance (opaque lines) and dissipation (semi-opaque lines) of device B at $B_\perp = 2.5$ T. The low dissipation confirms that the measurements are in the low-frequency limit, such that the dips in capacitance can be interpreted as corresponding to incompressible states. **c**, Conductance under the same conditions. The absence of a detectable change in capacitance, even as the two-terminal conductance undergoes a transition from an insulating state to a metallic state (Extended Data Fig. 6), suggests that the conductance transition is due to the emergence of gapless edge states.

device plane, held constant, the initially low charge-neutrality point conductance (G_{cnp}) increases steadily before finally saturating at $G \approx 1.8e^2/h$ for the largest total field applied (e, electron charge; h, Planck's constant). Evidence for a similar transition was recently reported in bilayer graphene[24], where the additional orbital degeneracy of the zLL leads to a conductance of $4e^2/h$. We note that although superficially similar, the structure and transport properties of the resulting edge modes are likely to be heavily influenced by the additional degeneracy, particularly when many-body reconstructions of the edge states are taken into account[10,25].

To distinguish the roles of the edges and the bulk in this conductance transition, we also measure the capacitance between the graphene and the graphite back gate under similar conditions. Capacitance (C) measurements serve as a probe of the bulk density of states (D) via $C^{-1} = C_G^{-1} + (Ae^2D)^{-1}$, where C_G is the geometric capacitance and A is the sample area. Simultaneous capacitance and transport measurements from a second graphene device show that quantized Hall states within the zLL at $\nu = 0$ and $\nu = \pm 1$ are associated with minima in the density of states (Fig. 1b, c). As the total field is increased, the capacitance dip at $\nu = 0$ remains unaltered even as the conductance increases

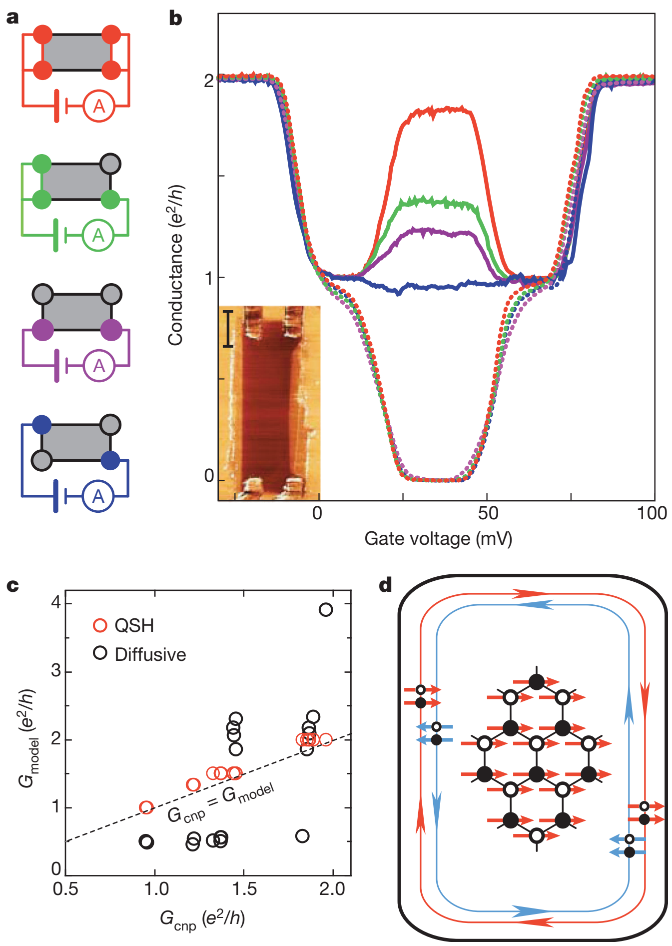


Figure 2 | Nonlocal two-terminal transport in the QSH regime. **a**, Schematic diagram of four distinct two-terminal measurement topologies available in a four-terminal device. Open circles indicate floating contacts whereas filled, coloured circles indicate measurement contacts. Each variation probes two parallel conductance paths between the measurement contacts with a variable number of segments on each path, indicated by black edges. **b**, Two-terminal conductance measurements of device A for $B_\perp = 1.4$ T, colour-coded to match the four different measurement configurations. Dashed curves correspond to $B_T = 1.4$ T; solid curves correspond to $B_T = 34.5$ T (QSH regime). In the QSH regime, G_{cnp} depends strongly on the number of floating contacts (see Extended Data Fig. 4 for similar data for device C). Inset, atomic force microscope (AFM) phase micrograph of device A; scale bar, 1 μm. **c**, G_{cnp} for eighteen different contact configurations based on cyclic permutations of the topologies shown in **a**. Data are plotted against two model fits. In a numerical simulation based on a diffusive model (black circles), the graphene flake was assumed to be a bulk conductor with the conductivity left as a fitting parameter ($\sigma = 3.25e^2/h$ for the best fit). The QSH model (red circles) is equation (1) and has no fitting parameters. The dashed line indicates a perfect fit of data to model. We note that the measured G_{cnp} never reaches the value predicted by the QSH model, indicating either contact resistance or finite backscattering between the helical edge states. **d**, Schematic diagram of bulk order and edge-state spin texture in the fully polarized QSH regime. Arrows indicate the projection of the electron spin on a particular sublattice, with the two sublattices indicated by open and filled circles. The edge-state wavefunctions are evenly distributed on the two sublattices and have opposite spin polarizations, at least for an idealized armchair edge[14].

by several orders of magnitude. This implies that the high-field $\nu = 0$ state has an incompressible bulk, consistent with the hypothesis of a ferromagnetic QSH state with conducting edge states and a bulk gap.

We probe the nature of the edge states through non-local transport measurements in which floating contacts are added along the sample edges[26]. Unlike the chiral edge of a quantum Hall state, which carries current in only one direction, the QSH edge can carry current either way, with backscattering suppressed by the conservation of spin within the helical edge states. Because the carriers do not maintain their spin coherence within a metal contact, contacts equilibrate the counterpropagating states such that each length of QSH edge between contacts must be considered a single resistor of resistance h/e^2. The two-terminal conductance results from the parallel addition of the two edges connecting the measurement probes:

$$G = \frac{e^2}{h}\left(\frac{1}{N_1+1} + \frac{1}{N_2+1}\right) \qquad (1)$$

Here N_1 and N_2 are the respective numbers of floating contacts along each edge. Figure 2b shows the results of non-local two-terminal conductance measurements for the four distinct two-terminal measurement geometries available in a four-terminal device (Fig. 2a). Repeating the measurement for 18 cyclic permutations of the available contact configurations, we find that the results are well fitted by the simple model of equation (1) (Fig. 2c), despite large variations in the effective bulk aspect ratio. Notably, G_{cnp} is always less than the value expected from the QSH model, suggesting some small but finite amount of backscattering or contact resistance. The combination of bulk incompressibility and non-local transport signatures of counterpropagating edge states leads us to conclude that the high-field metallic state observed indeed displays a QSH effect.

The QSH state realized here is equivalent to two copies of the quantum Hall effect, protected from mixing by the U(1) symmetry of spin rotations in the plane perpendicular to the magnetic field. As such, it constitutes a topologically non-trivial state that is clearly distinct in its edge-state properties from the insulating state at fully perpendicular field. Capacitance measurements in the intermediate conductance regime reveal that the bulk gap does not close as the total field is increased (Fig. 3a). This rules out a conventional topological phase transition, in which case the bulk gap is required to close[27]; the transition must thus occur by breaking the spin symmetry on which the QSH effect relies. In fact, a canted antiferromagnetic (CAF) state (Fig. 3b) that spontaneously breaks this symmetry is among the theoretically allowed $\nu = 0$ states[11–13]. In this scenario, the canting angle is controlled by the ratio of the Zeeman energy, $g\mu_B B_T$ ($g = 2$, bare gyromagnetic ratio; μ_B, Bohr magneton), and the antiferromagnetic exchange coupling, which depends only on $B_\perp$. The observed conductance transition results from the edge gap closing (Fig. 3c) as the spins on the two graphene sublattices are slowly canted by the in-plane magnetic field, with the fully polarized QSH state emerging above a critical value of B_T (ref. 14). In the language of SPT insulators, the antiferromagnetic instability breaks the spin symmetry below this critical field, allowing the counterpropagating edge states to backscatter and acquire a gap[28].

Experimentally, the subcritical field regime is characterized by high-conductance peaks appearing symmetrically between $\nu = 0$ and $\nu = \pm 1$. We observe $G > e^2/h$ peaks in many samples with widely varying aspect ratios (Extended Data Fig. 5), which is inconsistent with diffusive bulk transport in a compressible Landau level[29]. Measurements at different temperatures indicate that the peaks are metallic, even when the state at $\nu = 0$ is still strongly insulating (Fig. 4a). Moreover, the peaks exhibit the non-local transport behaviour of counterpropagating edge states (Fig. 4b); in particular, the peak conductance is always strictly less than e^2/h when the two edges are each interrupted by a floating contact. These results indicate that the conductance peaks are due to edge-state transport in the CAF state. The high conductance of these edge states, despite proximity to the strongly disordered etched graphene edge, implies that backscattering is at least partly suppressed. This is consistent with

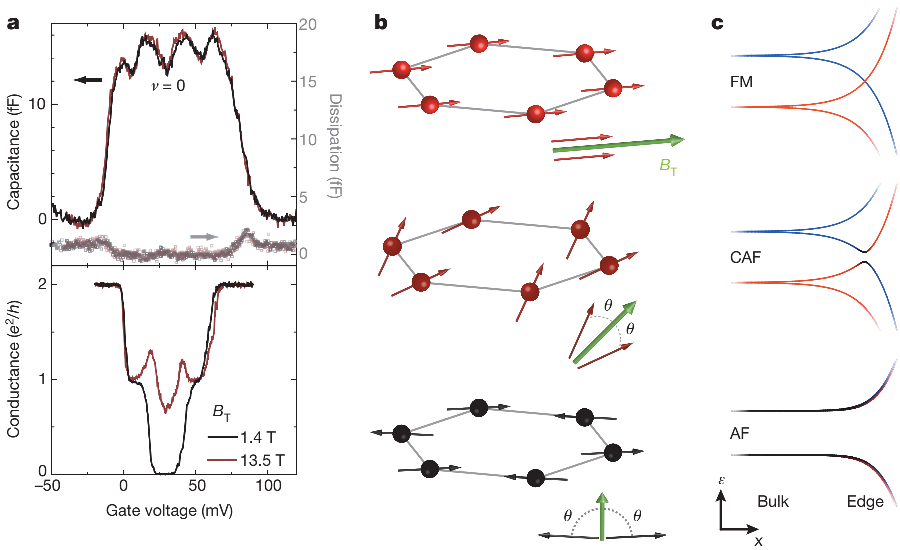


Figure 3 | Symmetry-driven quantum phase transition. **a**, Capacitance (top) and conductance (bottom) of device A at $B_\perp = 1.1$ T. The central dip in capacitance does not change with B_T at any point during the transition, implying that the bulk gap does not close. **b**, Bulk spin order in the three transition regimes. The balls and arrows are respectively schematic representations of the spin and sublattice textures of the ground-state wavefunctions and do not represent individual electrons; the electron density within the zLL at $\nu = 0$ is two electrons per cyclotron guiding centre. Insets, details of the relative alignment of the electron spins on the two sublattices. At large B_T, the bulk electron spins are aligned with the field (top panel), resulting in an emergent U(1) spin-rotation symmetry in the plane perpendicular to B_T. As the total magnetic field is reduced below some critical value (with $B_\perp$ held constant), the spins on opposite sublattices cant with respect to each other while maintaining a net polarization in the direction of B_T (middle panel). This state spontaneously breaks the U(1) symmetry, rendering local rotations of the electron spins energetically costly. For pure perpendicular fields (bottom panel), the valley isospin anisotropy energy overwhelms the Zeeman energy and the canting angle, θ, is close to 90°, defining a state with antiferromagnetic order. **c**, Low-energy band structure in the three phases[14]. ε is the energy and x is the in-plane coordinate perpendicular to the physical edge of the sample. The intermediate CAF phase smoothly interpolates between the gapless edge states of the QSH phase (top panel; FM, ferromagnetic) and the gapped edge of the perpendicular-field phase (bottom panel; AF, antiferromagnetic) without closing the bulk gap. Colour indicates the spin texture of the bands projected onto the magnetic field direction: red, aligned; blue, antialigned; black, zero net spin along the field direction.

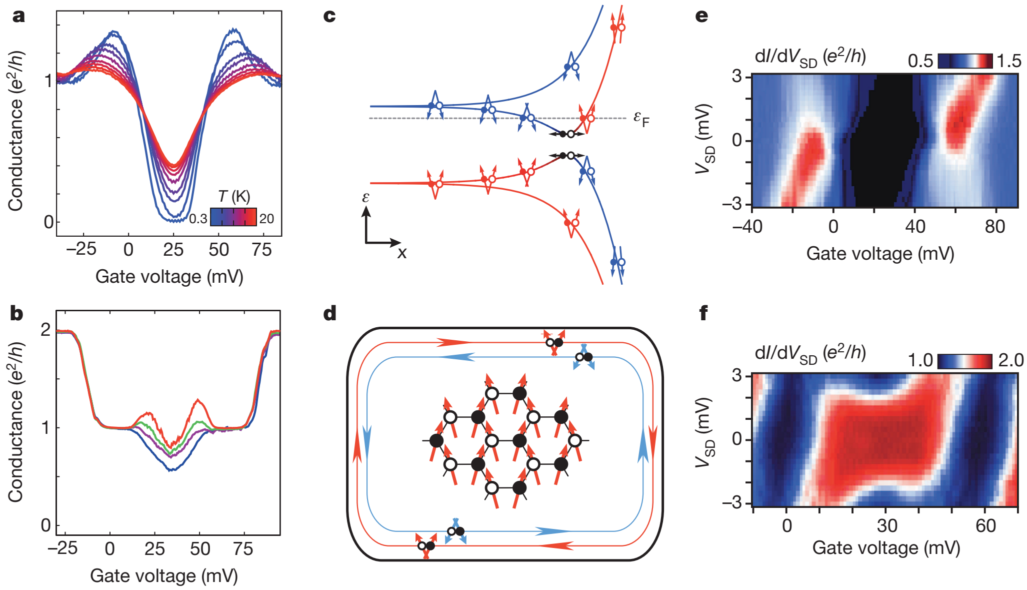


Figure 4 | Spin-textured edge states of the CAF phase. **a**, Temperature dependence in the intermediate-field regime for device C at $B_\perp = 5.9$ T and $B_T = 45.0$ T. The conductance peaks shows a metallic temperature coefficient, whereas the state at charge neutrality remains insulating. **b**, Non-local two-terminal conductance of device A at $B_\perp = 1.6$ T and $B_T = 26.1$ T. Colour-coding indicates contact geometry following the scheme in Fig. 2a. The height of the conductance peaks depends strongly on the configuration of floating contacts, indicating their origin in the gapped, counterpropagating edge states of the CAF phase. **c**, Schematic band diagram, including spin order, of the CAF edge states. For the electron and hole bands nearest to zero energy, the canting angle inverts near the sample edge, leading to counterpropagating edge states with inverted CAF spin texture. The dashed grey line indicates the Fermi energy, ε_F, in the regime corresponding to one of the conductance peaks. **d**, Schematic of bulk order and edge-state spin texture in the CAF regime, following the convention of Fig. 2d. **e**, Differential conductance, dI/dV_{SD}, of device C in the CAF regime ($B_\perp = 5.9$ T, $B_T = 45.0$ T) in units of e^2/h. A constant source–drain voltage, V_{SD}, along with a 100-μV, 313-Hz excitation voltage, are applied to one contact and the a.c. current is measured through the second, grounded contact. **f**, dI/dV_{SD} of device C in the QSH regime ($B_\perp = 2.7$ T, $B_T = 45.0$ T) in units of e^2/h. In both **e** and **f**, a symmetry is observed on reversing both V_{SD} and carrier polarity.

the theory of the band structure of the CAF state[14] (Fig. 4c, d), in which the interpolation between the gapless QSH and gapped antiferromagnetic edge structures is achieved by means of a new kind of one-dimensional edge state in which counterpropagating modes have oppositely canted antiferromagnetic spin textures. Notably, existing theories of the CAF state are only rigorously applicable to the zero-carrier-density regime, in which case the CAF edge modes exist as excited states. The fact that we can access the CAF edge states by gating is somewhat surprising, because it implies that this spectrum is stable to small populations of the edge bands.

Questions remain about the precise nature of the QSH and CAF boundary modes. The measured G_{cnp} never reaches $2e^2/h$ even at the highest values of B_T, despite some of the devices showing a flat plateau around charge neutrality. Naively, backscattering within the QSH edge mode requires flipping an electron spin, for example by magnetic impurities, although such a process should be energetically unfavourable at high magnetic fields. More trivially, we cannot exclude that weakly conducting charge puddles connect the two edges (but not source and drain contacts), leading to backscattering across the bulk in the QSH regime. Spin–orbit effects may also play a part by spoiling the spin symmetry on which the helical edge states rely. Although the intrinsic spin–orbit coupling in graphene is thought to be weak[17], the helical states may be uniquely sensitive to spin relaxation. Alternatively, the large Rashba-type spin–orbit coupling induced in the graphene under the gold contacts[30] may contribute a QSH-specific contact resistance that lowers the plateau conductance. The effects of disordered edges on helical edge transport have also not been addressed by modern theoretical treatments.

Nonlinear-transport measurements provide some additional insight into the nature of backscattering in the edge states. In both the QSH regime and the CAF regime, the nonlinear-transport data are invariant under simultaneous inversion of the carrier density and source–drain bias, V_{SD} (Fig. 4e, f). The data thus respect charge conjugation symmetry within the graphene, possibly implying that the inelastic processes probed at large V_{SD} values are native to the electronic system. Notably, the nonlinear conductance is not invariant under reversal of source drain bias alone. We can understand this lack of symmetry as a natural consequence of dissipative edge transport in our system, where, in contrast to TRS topological insulators, the counterpropagating edge states can be spatially separated. Within this picture, reversing V_{SD} changes the current carried by the inner and outer counterpropagating edge states. If dissipation differs between the two states on a single edge and the two physical graphene edges are inequivalent, reversing V_{SD} can be expected to result in a different conductance.

The experiments presented in this Letter demonstrate the CAF–QSH crossover in monolayer graphene. In addition, they enable the study of QSH physics in a versatile material platform, enabling new experiments. Most importantly, the high-field graphene QSH system differs from the conventional TRS QSH state through the crucial role of interactions, which lead to the spontaneous breaking of spin symmetry that generates the gapped CAF edge states. We note that in this paper we have discussed experimental results in the context of mean-field treatments of interactions in the graphene zLL[13,14]. Crucially, this treatment neglects the potential of the spin-ferromagnetic (or CAF) order parameter to reconstruct near the sample boundary[10,12], possibly leading to a qualitative change in the nature of the edge charge carriers. These results should inspire more careful experimental and theoretical work, both to understand the true nature of the edge states and to use them as a building block for realizing novel quantum circuits.

METHODS SUMMARY

Stacks of graphene, hexagonal boron nitride and graphite layers were fabricated by a dry transfer process. The sample surface was cleaned by high-temperature annealing in a reducing atmosphere after each transfer step and again after patterning of contacts by standard electron-beam lithography techniques. Before measurement, residual debris from the fabrication process was swept off the graphene flake with

an AFM tip operated in contact mode, the evidence of which is visible in the AFM micrograph inset in Fig. 2b.

Conductance measurements were made using a ~300-Hz voltage bias, with root mean squared amplitude $V_{rms} = 100\,\mu V$. The sample was immersed in ^3He liquid at 300 mK for all measurements except those shown in Fig. 4a, where the temperature is indicated, and those shown in Fig. 3, which were made at 150 mK with the sample immersed in a ^3He–^4He mixture. The angle between the magnetic field and the graphene plane was controlled by a mechanical rotator. The alignment angle was determined using high-density Shubnikov–de Haas oscillations, ensuring reproducible alignment to better than 0.025° in the large-tilt-angle regime, $B_\perp \ll B_T$. For multi-terminal devices (A and C), all measurements are done between two pairs of contacts (Fig. 3a, top configuration) unless otherwise indicated.

To measure capacitance, we used a cryogenic amplifier based on a high-electron-mobility transistor to construct a low-temperature capacitance bridge-on-a-chip, in which a 78-kHz a.c. excitation on the graphene sample was balanced against a variable phase and amplitude excitation on a known reference capacitor. The sample bias was 900 μV for the data set in Fig. 1b and 100 μV for that in Fig. 3a. The signal at the input of the cryogenic amplifier was first nulled by adjusting the reference excitation using a dual-channel a.c. signal generator, after which data were acquired off-balance by monitoring the in-phase and out-of-phase voltages at the balance point as a function of the applied d.c. sample bias. Biasing the transistor amplifier raised the base temperature of the cryostat, such that the temperature was 400 mK during acquisition of the data in Fig. 1b and 250 mK during acquisition of that in Fig. 3a.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 July; accepted 22 October 2013.
Published online 22 December 2013.

1. Ryu, S., Schnyder, A. P., Furusaki, A. & Ludwig, A. W. W. Topological insulators and superconductors: tenfold way and dimensional hierarchy. *N. J. Phys.* **12,** 065010 (2010).
2. Chen, X., Gu, Z.-C., Liu, Z.-X. & Wen, X.-G. Symmetry-protected topological orders in interacting bosonic systems. *Science* **338,** 1604–1606 (2012).
3. Hasan, M. Z. & Kane, C. L. Topological insulators. *Rev. Mod. Phys.* **82,** 3045–3067 (2010).
4. Qi, X.-L. & Zhang, S.-C. Topological insulators and superconductors. *Rev. Mod. Phys.* **83,** 1057–1110 (2011).
5. Chen, Y. L. *et al.* Massive Dirac fermion on the surface of a magnetically doped topological insulator. *Science* **329,** 659–662 (2010).
6. Xu, S.-Y. *et al.* Hedgehog spin texture and Berry's phase tuning in a magnetic topological insulator. *Nature Phys.* **8,** 616–622 (2012).
7. König, M. *et al.* Quantum spin Hall insulator state in HgTe quantum wells. *Science* **318,** 766–770 (2007).
8. Du, L., Knez, I., Sullivan, G. & Du, R.-R. Observation of quantum spin Hall states in InAs/GaSb bilayers under broken time-reversal symmetry. Preprint at http://arxiv.org/abs/1306.1925 (2013).
9. Abanin, D. A., Lee, P. A. & Levitov, L. S. Spin-filtered edge states and quantum Hall effect in graphene. *Phys. Rev. Lett.* **96,** 176803 (2006).
10. Fertig, H. A. & Brey, L. Luttinger liquid at the edge of undoped graphene in a strong magnetic field. *Phys. Rev. Lett.* **97,** 116805 (2006).
11. Herbut, I. F. Theory of integer quantum Hall effect in graphene. *Phys. Rev. B* **75,** 165411 (2007).
12. Jung, J. & MacDonald, A. H. Theory of the magnetic-field-induced insulator in neutral graphene sheets. *Phys. Rev. B* **80,** 235417 (2009).
13. Kharitonov, M. Phase diagram for the $\nu = 0$ quantum Hall state in monolayer graphene. *Phys. Rev. B* **85,** 155439 (2012).
14. Kharitonov, M. Edge excitations of the canted antiferromagnetic phase of the $\nu = 0$ quantum Hall state in graphene: a simplified analysis. *Phys. Rev. B* **86,** 075450 (2012).
15. Hsieh, T. H. *et al.* Topological crystalline insulators in the SnTe material class. *Nature Commun.* **3,** 982 (2012).
16. Kindermann, M. Topological crystalline insulator phase in graphene multilayers. Preprint at http://arxiv.org/abs/1309.1667 (2013).
17. Min, H. *et al.* Intrinsic and Rashba spin-orbit interactions in graphene sheets. *Phys. Rev. B* **74,** 165310 (2006).
18. Semenoff, G. W. Condensed-matter simulation of a three-dimensional anomaly. *Phys. Rev. Lett.* **53,** 2449–2452 (1984).
19. Checkelsky, J. G., Li, L. & Ong, N. P. Zero-energy state in graphene in a high magnetic field. *Phys. Rev. Lett.* **100,** 206801 (2008).
20. Yang, K., Das Sarma, S. & MacDonald, A. H. Collective modes and skyrmion excitations in graphene SU(4) quantum Hall ferromagnets. *Phys. Rev. B* **74,** 075423 (2006).
21. Alicea, J. & Fisher, M. P. A. Graphene integer quantum Hall effect in the ferromagnetic and paramagnetic regimes. *Phys. Rev. B* **74,** 075422 (2006).
22. Nomura, K., Ryu, S. & Lee, D.-H. Field-induced Kosterlitz-Thouless transition in the $N = 0$ Landau level of graphene. *Phys. Rev. Lett.* **103,** 216801 (2009).
23. Young, A. F. *et al.* Spin and valley quantum Hall ferromagnetism in graphene. *Nature Phys.* **8,** 550–556 (2012).
24. Maher, P. *et al.* Evidence for a spin phase transition at charge neutrality in bilayer graphene. *Nature Phys.* **9,** 154–158 (2013).
25. Mazo, V., Huang, C.-W., Shimshoni, E., Carr, S. T. & Fertig, H. A. Superfluid-insulator transition of quantum hall domain walls in bilayer graphene. Preprint at http://arxiv.org/abs/1309.1563 (2013).
26. Roth, A. *et al.* Nonlocal transport in the quantum spin Hall state. *Science* **325,** 294–297 (2009).
27. Büttner, B. *et al.* Single valley Dirac fermions in zero-gap HgTe quantum wells. *Nature Phys.* **7,** 418–422 (2011).
28. Shitade, A. *et al.* Quantum spin Hall effect in a transition metal oxide Na_2IrO_3. *Phys. Rev. Lett.* **102,** 256403 (2009).
29. Abanin, D. A. & Levitov, L. S. Conformal invariance and shape-dependent conductance of graphene samples. *Phys. Rev. B* **78,** 035416 (2008).
30. Marchenko, D. *et al.* Giant Rashba splitting in graphene due to hybridization with gold. *Nature Commun.* **3,** 1232 (2012).

Acknowledgements We acknowledge discussions with D. Abanin, A. Akhmerov, C. Beenakker, L. Brey, L. Fu, M. Kharitonov, L. Levitov, P. Lee and J. Sau. B.H. and R.C.A. were funded by the BES Program of the Office of Science of the US DOE, contract no. FG02-08ER46514, and the Gordon and Betty Moore Foundation, through grant GBMF2931. J.D.S.-Y, and P.J.-H. were primarily supported by the US DOE, BES Office, Division of Materials Sciences and Engineering, under award DE-SC0001819. Early fabrication feasibility studies were supported by NSF Career Award no. DMR-0845287 and the ONR GATE MURI. This work made use of the MRSEC Shared Experimental Facilities supported by the NSF under award no. DMR-0819762 and of Harvard's CNS, supported by the NSF under grant no. ECS-0335765. Some measurements were performed at the National High Magnetic Field Laboratory, which is supported by NSF Cooperative Agreement DMR-0654118, the State of Florida and the DOE. A.F.Y. acknowledges the support of the Pappalardo Fellowship in Physics.

Author Contributions A.F.Y. and J.D.S.-Y. had the idea for the experiment. J.D.S.-Y. and S.H.C. fabricated the samples. A.F.Y., J.D.S.-Y. and B.H. performed the experiments, analysed the data and wrote the paper. T.T. and K.W. grew the crystals of hexagonal boron nitride. R.C.A. and P.J.-H. advised on experiments, data analysis and writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.F.Y. (afy@mit.edu), J.D.S.-Y. (jdsy@mit.edu) or B.H. (benhunt@mit.edu).

LETTER

doi:10.1038/nature12780

Dislocations in bilayer graphene

Benjamin Butz[1], Christian Dolle[1], Florian Niekiel[1], Konstantin Weber[2], Daniel Waldmann[3], Heiko B. Weber[3], Bernd Meyer[2] & Erdmann Spiecker[1]

Dislocations represent one of the most fascinating and fundamental concepts in materials science[1–3]. Most importantly, dislocations are the main carriers of plastic deformation in crystalline materials[4–6]. Furthermore, they can strongly affect the local electronic and optical properties of semiconductors and ionic crystals[7,8]. In materials with small dimensions, they experience extensive image forces, which attract them to the surface to release strain energy[9]. However, in layered crystals such as graphite, dislocation movement is mainly restricted to the basal plane. Thus, the dislocations cannot escape, enabling their confinement in crystals as thin as only two monolayers. To explore the nature of dislocations under such extreme boundary conditions, the material of choice is bilayer graphene, the thinnest possible quasi-two-dimensional crystal in which such linear defects can be confined. Homogeneous and robust graphene membranes derived from high-quality epitaxial graphene on silicon carbide[10] provide an ideal platform for their investigation. Here we report the direct observation of basal-plane dislocations in freestanding bilayer graphene using transmission electron microscopy and their detailed investigation by diffraction contrast analysis and atomistic simulations. Our investigation reveals two striking size effects. First, the absence of stacking-fault energy, a unique property of bilayer graphene, leads to a characteristic dislocation pattern that corresponds to an alternating AB↔AC change of the stacking order. Second, our experiments in combination with atomistic simulations reveal a pronounced buckling of the bilayer graphene membrane that results directly from accommodation of strain. In fact, the buckling changes the strain state of the bilayer graphene and is of key importance for its electronic properties[11–14]. Our findings will contribute to the understanding of dislocations and of their role in the structural, mechanical and electronic properties of bilayer and few-layer graphene.

In graphite, the bulk material most closely related to graphene, basal-plane dislocations have been well known since the early 1960s[15]. The dissociation of perfect basal-plane dislocations into pairs of Shockley partial dislocations (partials) bounding a stacking-fault ribbon in the basal plane was observed using conventional transmission electron microscopy[15,16] (TEM). The separation of partials (or the width of the stacking-fault ribbon) is determined by the balance between the repulsive forces (reduction of total strain energy) and the attractive forces (minimization of stacking-fault energy). Similarly, the dissociation of perfect dislocations into Shockley partials plays an essential role for most materials with face-centred-cubic structure and is of particular importance for the plasticity of face-centred-cubic metals[9].

Also in few-layer graphene, changes of the stacking sequence have been reported[17–22]. For instance, dark-field TEM was used to reveal changes in the stacking order and twisting of freestanding bi- and trilayer graphene grown by chemical vapour deposition[18] (CVD). However, the partials necessarily connected to the stacking faults were not addressed in that study. Moreover, scanning tunnelling microscopy was used to investigate local changes in the stacking sequence of few-layer graphene on mica[19]. The boundaries, which separate areas of different stacking order, form a pattern that closely resembles dislocation networks in bulk graphite, indicating that it may be possible to study basal-plane dislocations in few-layer graphene. Two recent publications[17,20] address in more detail the local stacking transition AB↔AC in bilayer graphene grown by CVD (subsequently stacked[17] or as-grown[20]). In both studies, the authors applied dark-field TEM to investigate the local bilayer graphene stacking and to characterize the distribution of the transition regions. Moreover, atomistic models were derived from comparison with high-resolution scanning TEM (STEM) images. The transition regions were described either in terms of strain solitons using a Frenkel–Kontorova model[17] or with complicated atomistic models, including the delamination of the two graphene layers, which were proposed, on the basis of molecular dynamics simulation, to explain the experimental data[20]. However, none of this work describes the transition regions in terms of classical dislocation theory.

Important modifications of dislocations are expected when going from graphite to freestanding few-layer graphene membranes. First, pronounced buckling of the membrane is expected as a result of strain accommodation. Moreover, in bilayer graphene the unique situation occurs that both stacking variants, AB and AC (or BA, which means the same), are equivalent. This means that no stacking-fault energy exists and, therefore, that no attractive forces act between partials to reduce the stacking-fault area.

In the following, we present the detailed microscopic study of basal-plane dislocations in bilayer graphene membranes prepared from high-quality epitaxial graphene on the (0001) surface of 6H-SiC (ref. 10; throughout the manuscript four Bravais–Miller indices are used to account for the hexagonal symmetry). The central region of a typical TEM sample with more than 150 membranes is shown in Fig. 1a. The colour-coded STEM image of one membrane (Fig. 1b) indicates the local variation in the number of graphene layers as determined by Raman microscopy[10] and TEM (see below). Apart from the bilayer

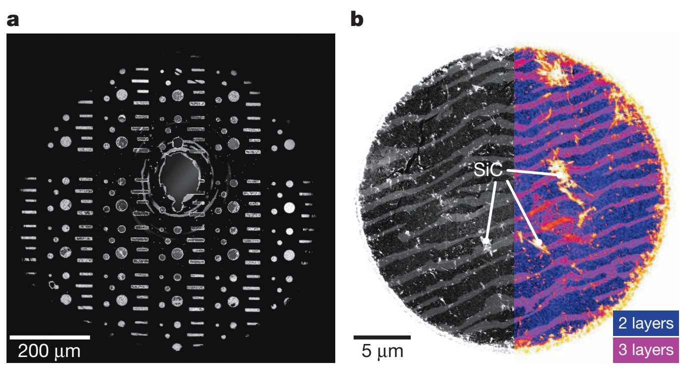


Figure 1 | Freestanding bilayer graphene membranes in a SiC frame. **a**, Dark-field STEM image (obtained in a scanning electron microscope) of a sample with circular and rectangular membranes. The non-transparent SiC is dark and the freestanding membranes appear bright. **b**, Single membrane at higher magnification. The colour code indicates the local number of graphene layers (compare with ref. 10).

[1]Center for Nanoanalysis and Electron Microscopy, Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstraße 6, 91058 Erlangen, Germany. [2]Interdisziplinäres Zentrum für Molekulare Materialien und Computer-Chemie-Centrum, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstraße 25, 91052 Erlangen, Germany. [3]Lehrstuhl für Angewandte Physik, Friedrich-Alexander-Universität Erlangen-Nürnberg, Staudtstraße 7, 91058 Erlangen, Germany.

graphene, the membranes exhibit regions consisting of three (or even four) graphene layers.

To study the microstructure and crystal defects in the graphene membrane, we employed dark-field TEM imaging. Figure 2a shows a $\{11\bar{2}0\}$ dark-field image of a typical membrane area with predominantly bilayer graphene and a smaller portion of trilayer graphene. The bilayer and trilayer nature of the corresponding areas is confirmed by a detailed rocking-curve analysis based on a $\{11\bar{2}0\}$ dark-field tilt series (Fig. 2c, Extended Data Fig. 2 and Supplementary Video 1).

In the following, we concentrate on the bilayer region. Most notably, a dense network of dislocations, confined between the two graphene sheets, appears as relatively sharp dark lines in Fig. 2a. As confirmed by the detailed Burgers vector analysis (see below), all the dislocations are partials, meaning that their Burgers vectors are of type $b = (1/3)\langle 1\bar{1}00\rangle$ and therefore do not correspond to lattice translations in the basal plane. Thus, each partial is associated with a change of the local stacking between AB and AC, or vice versa. This is shown in the corresponding $(2\bar{2}00)$ dark-field image in Fig. 2b, where the two stacking orders in the bilayer regions appear with different intensities.

All this is congruent with dislocation theory in materials science, where stacking faults in three-dimensional (3D) crystals are always bordered by partials. However, the essential difference is that the stacking-fault energy in bilayer graphene is zero because AB and AC are equivalent by symmetry; in other words, a non-zero stacking-fault energy arises only if a third graphene layer is added. In line with this consideration, Fig. 2b shows that the two stacking variants occupy about equal portions of the bilayer area (in contrast to the trilayer region). Single partials in trilayer graphene directly cause a change in the crystal structure, for example from Bernal stacking to rhombohedral stacking (ABA $\leftrightarrow$ ABC), resulting in real stacking faults[17–22] (compare the respective trilayer regions in Figs 2b and 3b). Apart from regions with an apparently irregular distribution of dislocations, for example close to the trilayer region in Fig. 2a, we found that a large fraction of our bilayer graphene exhibited parallel aligned partials (see, for example, the centre of Fig. 2a). The typical spacing between individual partials in such regions is in the range of 20–30 nm. Concerning the origin of the dislocations, TEM studies of membranes covering very small substrate holes (Extended Data Fig. 3) indicate that a large portion of the dislocations must already be present in the graphene on the SiC wafer before the substrate is removed. We assume that the dislocations form as result of misfit stresses either during the growth of the graphene on the SiC or while the samples are cooled from 1,750 °C to room temperature (owing to different thermal expansion coefficients).

The central region of Fig. 2a with the characteristic equidistant arrangement of parallel dislocations was selected to analyse the Burgers vectors and strain fields of the partials in detail. By systematic dark-field imaging with specific $\{11\bar{2}0\}$ and $\{1\bar{1}00\}$ reflections, the dislocations are unambiguously identified as 60° partials with alternating Burgers vectors of type $b = (1/3)\langle 1\bar{1}00\rangle$ (Fig. 3). This result matches the alteration of the stacking sequence (AB $\leftrightarrow$ AC $\leftrightarrow$ AB) as observed in Fig. 2b. As a consequence, each pair of partials (in the homogeneous bilayer regions) results in a perfect lattice translation along $\langle 11\bar{2}0\rangle$ corresponding to an effective perfect edge dislocation.

To explore the peculiarities of basal-plane dislocations in bilayer graphene, we carried out atomistic simulations of the equidistant arrangement of partials. Starting with a perfect edge dislocation, splitting into equidistant 60° partials occurs during structural optimization (Fig. 4a). This is in agreement both with our experimental observation and with energetic arguments based on the absence of stacking-fault energy. Apart from the alteration of the stacking (Fig. 4a, top view), the most notable feature is the pronounced buckling (by around 1 nm) of the membrane (Fig. 4a, 3D visualization and side view). Unlike the well-known intrinsic corrugations of perfect monolayer graphene[23], this buckling is directly associated with the dislocation lines because the driving force for the buckling is release of strain energy in the dislocation strain field. To understand the impact of buckling on the strain state, we compare the buckled configuration (Fig. 4a) with a flat one, where the movement of the atoms during the atomic relaxation was restricted to the x–y plane. The latter configuration can be thought to represent bilayer graphene on a flat compliant substrate, for example a wafer with an appropriate amorphous surface layer.

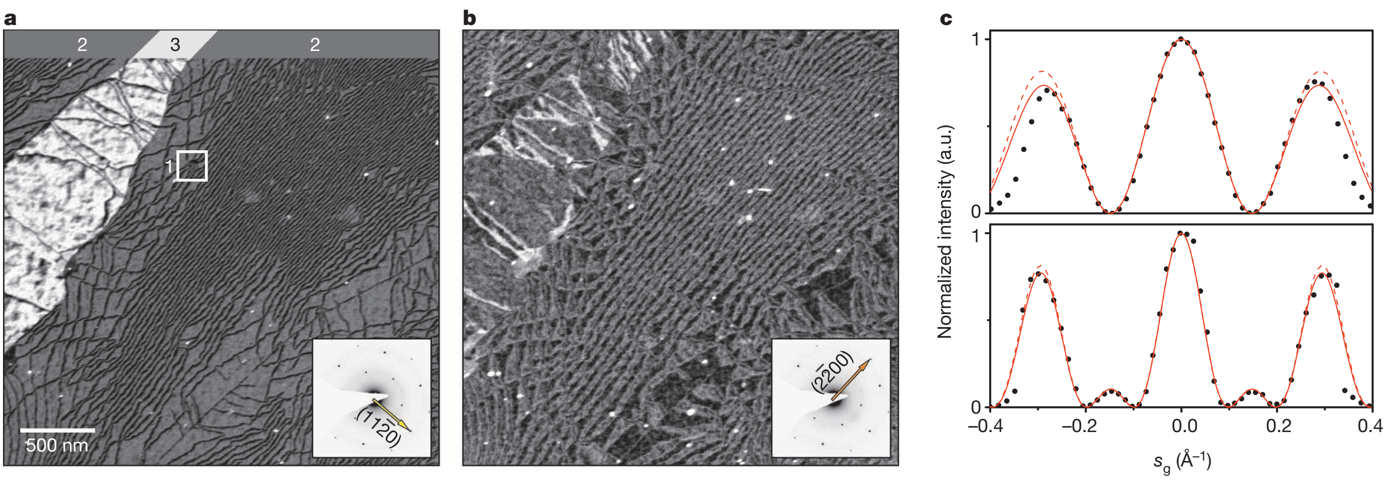


Figure 2 | Microstructure analysis by dark-field TEM imaging. a, b, Dark-field TEM images of the same area of a membrane obtained with the $(11\bar{2}0)$ and $(2\bar{2}00)$ reflections. The diffraction patterns indicate the active reflection for dark-field imaging (SEM overview and respective bright-field images are given in Extended Data Fig. 1). The $(11\bar{2}0)$ dark-field image (**a**) reveals bilayer and trilayer graphene as indicated by the numbers. The dark lines in the bilayer/trilayer regions correspond to an extended network of partials confined between the graphene layers (see Fig. 3 for details). The line contrast results from the altered diffraction in the strain/displacement field of the dislocations. Because the $\{11\bar{2}0\}$ reflections show identical structure factors (and rocking curves) for both stacking variants, AB and AC[18], the stacking faults bordered by the partials are invisible irrespective of the sample tilt. The local transition of the stacking sequence in the bilayer graphene from AB to AC, and vice versa, caused by partials observed in **a**, is clearly visible as areal intensity variation in the $(2\bar{2}00)$ dark-field image (**b**). Unlike the $\{11\bar{2}0\}$ reflections in **a**, $\{1\bar{1}00\}$ and $\{2\bar{2}00\}$ reflections show different rocking curves for AB and AC stacking[18], meaning that small deviations from the Bragg condition lead to stacking-fault contrast as seen in the image. **c**, Rocking curves of $\{11\bar{2}0\}$ dark-field intensities (in arbitrary units, a.u.) in dislocation-free areas of bi- and trilayer graphene (black dots) confirming the local number of graphene layers (see Extended Data Fig. 2 for details). The red lines are calculated rocking curves for bi- and trilayer graphene (equilibrium interlayer distance of 3.370 Å) taking the atomic scattering factor of carbon (dashed line) plus thermal damping (solid line) into account (see Extended Data Fig. 2 for details). The abscissa shows the excitation error s_g, which is related to the tilt angle and thus characterizes the scattering condition for each respective dark-field image.

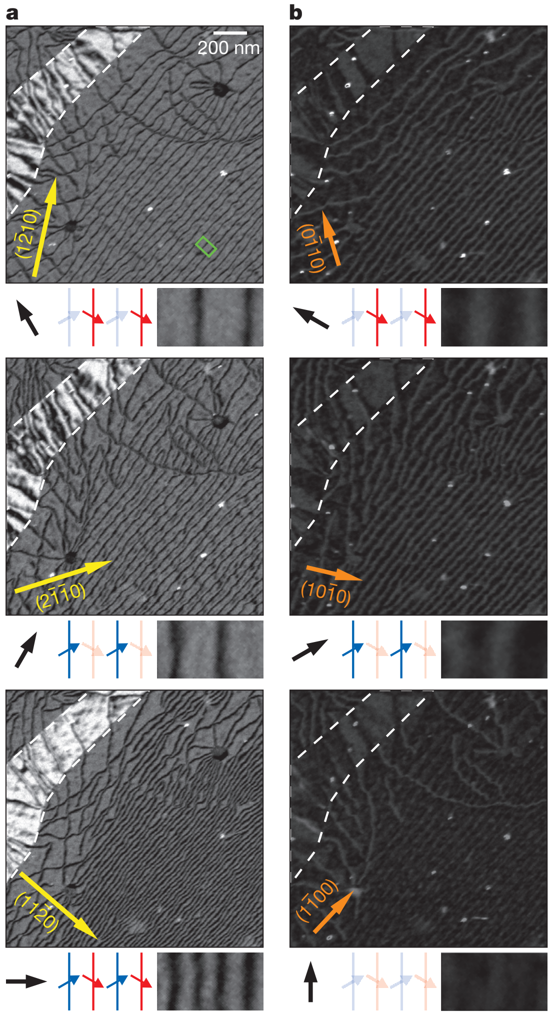


Figure 3 | Burgers vector analysis. **a**, **b**, Series of $\{11\bar{2}0\}$ (**a**) and $\{1\bar{1}00\}$ (**b**) dark-field images of the same region as in Fig. 2a (the corresponding diffraction pattern is given in Extended Data Fig. 4). The $\{11\bar{2}0\}$ reflections (**a**) have a large structure factor for both AB and AC stacking[18] owing to the constructive interference for all atoms, such that relative shifts of the two graphene sheets in the displacement field of the dislocation lowers the local diffracted intensity. Therefore, the dislocations appear as dark lines on a bright background. In contrast, for $\{1\bar{1}00\}$ dark-field images (**b**), they appear as bright lines on a dark background because the structure factor for such reflections is small owing to partial destructive interference of scattering waves from atoms in the A and B planes, as a result of which enhanced intensity is obtained in the displacement field of the dislocations. In contrast to $\{2\bar{2}00\}$ dark-field images (Fig. 2b), the $\{1\bar{1}00\}$ dark-field images do not show pronounced stacking-fault contrast because the deviation from the Bragg condition is too small for normal incidence. The diagram below each image shows the Burgers vector analysis, that is, a comparison of $\{11\bar{2}0\}$ dark-field images of the same parallel dislocations (green box and magnified sections below the images). In the first two images only every second dislocation is visible (the ones that are seen in the first image are absent from the second one, and vice versa), whereas the third image shows both. Left of the enlarged dark-field images are schematic diagrams of the orientation relation between the respective diffraction vectors g (black arrow) and Burgers vectors b (red and blue arrows). By applying the $g \cdot b = 0$ invisibility criterion, the two dislocation types are identified as 60° partials with Burgers vectors $b_1 = (1/3)[10\bar{1}0]$ (blue) and $b_2 = (1/3)[01\bar{1}0]$ (red). On the basis of these Burgers vectors, the contrast of the $\{1\bar{1}00\}$ dark-field images (**b**) is understood as well. Dislocations show up with relatively strong contrast if g and b are parallel or antiparallel ($|g \cdot b| = 2/3$), whereas they almost vanish if the angle between g and b is 60° ($|g \cdot b| = 1/3$). Applying this finding to our reference region, only one variant of the partials shows up in the top two $\{1\bar{1}00\}$ dark-field images, whereas only faint or residual contrast is seen in the third dark-field image. In **b**, the areal contrast variations in the trilayer region, marked by dashed lines, are due to changes of the stacking sequence from ABA (bright) to ABC (dark).

Before discussing the phenomenon of strain accommodation in detail, we make comparison with the TEM experiments to validate the model and to confirm the Burgers vector analysis. In this relation, dark-field images were calculated using both atomistic configurations (buckled and flat). Figure 4b compares the experimental and simulated dislocation contrast in $\{11\bar{2}0\}$ dark-field images (other reflections are shown in Extended Data Fig. 4). Within the accuracy of the experiment (owing to residuals on the membrane, the contrast varies slightly between the dislocations), there is very good agreement between experiment and simulation (especially using the buckled atomic configuration; see Fig. 4b). In particular, the applicability of the $g \cdot b = 0$ invisibility criterion for determination of the Burgers vector is confirmed.

Most remarkably, the dislocation-induced buckling of the bilayer graphene could be validated by comparing the second-order derivative calculated from experimental images (Fig. 4c, lower part) with the respective distributions obtained from both simulations (Fig. 4c, upper part). The same curvature (given by the second derivative) at the intensity maxima and minima of the original $(11\bar{2}0)$ dark-field image, reflected in bright lines with similar intensities in the modulus of the second derivative, is an excellent fit to the prediction of our simulation with the buckled atomic configuration. Prospectively, membranes with improved quality will allow for a more precise evaluation of the intensity distribution across single dislocations, enabling both the optimization of the theoretical interaction potential between graphene layers and the determination of the Peierls potential of the dislocations. These results will help to explain the mechanical properties of bilayer and few-layer graphene.

Figure 4d and Fig. 4e compare the respective profiles of the in-plane strain components, ε_{xx}, ε_{yy} and ε_{xy}, for the two layers in both configurations. It can be seen that the buckling completely alters the strain state of the two graphene layers. In the flat configuration, the normal strain component, ε_{xx}, shows pronounced maxima and minima at the dislocations cores (localized tension and compression of the graphene layers), whereas the same strain component almost completely levels out on buckling, resulting in a small and uniform strain (tensile in the lower layer but compressive in the upper layer) across the whole bilayer graphene ribbon. For symmetry reasons, ε_{yy} is almost zero for the two layers (in both configurations). Unexpectedly, the shear component, ε_{xy}, is much more localized in the buckled configuration. As a result, both the local disregistry, that is, the relative shift of the two layers, and the Burgers vector distributions are completely different for the buckled and the flat configurations (compare the respective distributions in the bottom parts of Fig. 4d and Fig. 4e).

Our simulations demonstrate how sensitively the strain state and the local stacking of the two graphene layers depend on the topography of the bilayer graphene. This is of great interest because one key aim of the field is to engineer a suitable bandgap for applications using strained bilayer or few-layer graphene. However, the dislocation-induced local buckling and, in particular, the resulting strain redistribution have been neglected in the literature so far[14,17,20].

It is worth comparing the basal-plane dislocations in bilayer graphene with dislocations in monolayer graphene (see, for example, refs 24–26) and putting both in the context of classical dislocation theory. At first glance, the two types of dislocations seem fundamentally different. Whereas the former lie in the plane of the membrane and can be as long as several micrometres, the dislocations in monolayer graphene possess an infinitesimally short dislocation line perpendicular to the graphene membrane (making even the definition of a dislocation line questionable) and thus seem more like topological defects in a two-dimensional (2D) crystal. However, in both cases the membranes are embedded in 3D space, allowing for strain relief in the third dimension by, for example, buckling (see, for example, refs 25, 27 for monolayer graphene). In fact, dislocations intersecting a thin plate in the normal direction were studied in elasticity theory more than 60 years ago[28]. The focus was on screw dislocations (for which there is no equivalent in monolayer graphene), but in relation to edge dislocations it was stated that in "certain circumstances an edge dislocation will be able to relieve most of its stress by slight

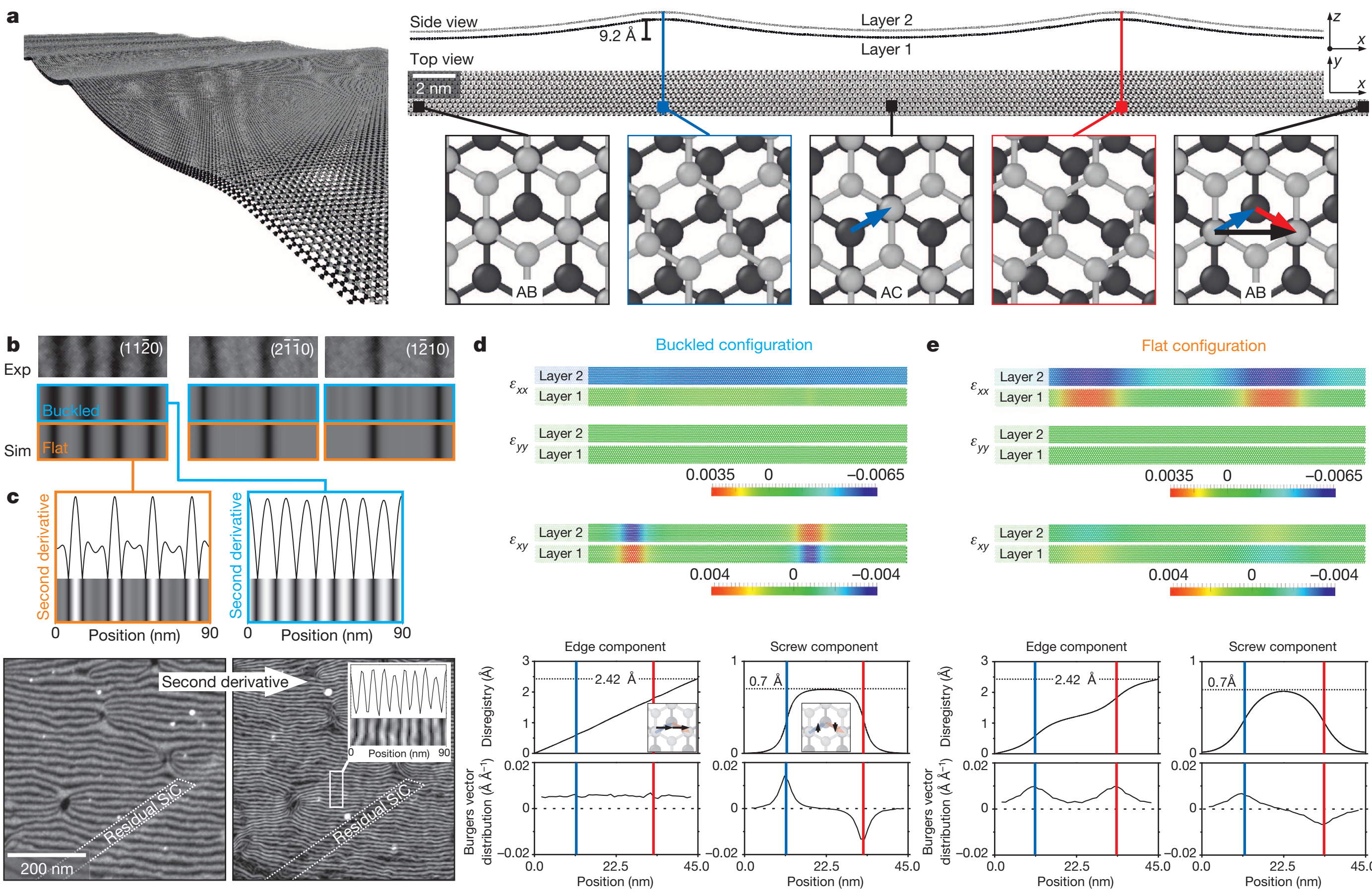


Figure 4 | Atomistic simulation of 60° partial dislocations: atomic configuration, TEM simulation, strain distribution and Burgers vector distribution. **a**, Atomistic simulation of a pair of parallel 60° partials as observed in our bilayer graphene (compare with Fig. 3): 3D visualization, side view, top view. The blue and red lines mark the dislocation cores. The gradual change of the stacking sequence, AB ↔ AC ↔ AB, across the two partials is shown enlarged together with the corresponding Burgers vectors of type $(1/3)\langle 1\bar{1}00\rangle$ (blue and red). The total relative shift (black vector) amounts to one lattice translation along the x direction (total Burgers vector of type $(1/3)\langle 11\bar{2}0\rangle$; for example, $(1/3)[10\bar{1}0] + (1/3)[01\bar{1}0] = (1/3)[11\bar{2}0]$). The local bond lengths and angles are depicted in Extended Data Fig. 6. **b**, Comparison of experimental $\{11\bar{2}0\}$ dark-field images (compare with Fig. 3) with simulated ones that are based on the buckled configuration in **a** (blue outline) and, as reference, a flat one (orange outline). The specific strain states of these atomic configurations cause different intensity distributions across the dislocations in dark-field imaging. In particular, the widths of the respective intensity minima are remarkably different for the two configurations, most notably in the $(11\bar{2}0)$ dark-field images, in which the edge components of both dislocations are directly probed (further dark-field images with other reflections are depicted in Extended Data Fig. 4). **c**, Proof of local buckling using the second derivative of the dark-field intensity distributions. The second-derivative profiles (obtained from **b**), the absolute values of which are plotted in the upper part, are distinctively different for the buckled and the flat configurations. Applying the same procedure to experimental $(11\bar{2}0)$ dark-field images (lower part) validates our buckled atomistic configuration. The inset in the second derivative of the experimental image depicts a representative intensity profile (averaging over 3 pixels) with characteristic variations. **d**, **e**, Atomistic strain components, ε_{xx}, ε_{xy} and ε_{yy}, for the two layers of the buckled configuration (**d**) and for those of the flat reference configuration (**e**), where the movement of the atoms during optimization was restricted to the x–y plane (fixed z coordinate). The distributions are shown together with the disregistry, that is, the relative shift of the two graphene layers, and the Burgers vector distributions for the edge and screw components of the partials.

buckling of the plate”. Along similar lines, the buckling of a thin foil due to dislocations lying in the plane of the foil (such as our basal plane dislocations in bilayer graphene) has also been investigated[29]. As in our case, the buckling is caused by an in-plane edge component of the Burgers vector. Thus, both the edge dislocations in monolayer graphene and the basal-plane dislocations in bilayer graphene can be viewed as limiting cases of classical dislocations in thin plates.

We have studied basal-plane dislocations in freestanding bilayer graphene—the thinnest possible crystal that can host such dislocations—by combining dark-field TEM and atomistic simulations. In contrast to dislocations in monolayer graphene, the dislocations reported here are real line defects confined between the two graphene layers. By applying Burgers vector analysis, we unambiguously identify the dislocations as partial dislocations with $b = (1/3)\langle 1\bar{1}00\rangle$, causing a change of the local stacking from AB to AC, and vice versa. The absence of stacking-fault energy, a unique peculiarity of bilayer graphene, gives rise to a characteristic equidistant arrangement of dislocations with alternating Burgers vectors observed in large sample areas. One outcome of this study is that pronounced buckling of the bilayer graphene membrane at the dislocations enables the partial compensation of the normal strain and, surprisingly, the complete delocalization of the respective residual compressive and tensile strains in the two graphene layers. This makes the dislocations in freestanding bilayer graphene distinctly different from corresponding ones in graphite or other 3D crystals. In contrast to recent publications on strain solitons[17] and stacking-fault boundaries[20] based on observations in graphene grown by CVD, in our treatment of the investigated one-dimensional topological defects in bilayer graphene we consistently use the well-established concept of dislocations in crystalline solids and extend it to quasi-2D crystals.

We expect our findings to contribute to our understanding of basal-plane dislocations and their role in tailoring the mechanical and electronic properties of bilayer and few-layer graphene. Our observation that such dislocations are already present in the initial epitaxial graphene on SiC will help explicate the restrictions on the transport properties of

this high-quality graphene material and should stimulate further studies on the local defect structure and related changes in the electronic properties. Furthermore, we anticipate that our freestanding bilayer graphene membranes with a well-defined distribution of parallel partial dislocations are good candidates for fundamental studies on (anisotropic) electronic transport in bilayer graphene and its potential use in future applications.

METHODS SUMMARY

The investigated graphene membranes were prepared from high-quality epitaxial graphene on 6H-SiC according to a procedure recently described[10]. State-of-the-art scanning electron microscopy (SEM) and aberration-corrected TEM including electron diffraction were used to investigate the local microstructure of the membranes, for example the distribution of graphene layers and their local stacking, as well as the extended network of basal-plane dislocations in the bilayer graphene. To confirm the local number of graphene layers, rocking curves of respective membrane regions (bi-, tri- and four-layer graphene) were determined from $\{11\bar{2}0\}$ dark-field tilt series. In particular, the specific properties of the observed partial dislocations, such as Burgers vector, dislocation type and strain-field distribution, were examined in detail by a systematic dark-field analysis. To prove the local buckling (expected from our simulations) as well as the associated strain redistribution at the dislocation cores, the experimental data were quantitatively compared to simulated dark-field TEM images. Those TEM image simulations were based on atomic configurations of the partials, which we derived from calculations using atomistic interatomic potentials. To reproduce the periodic arrangement of alternating partials as observed in our bilayer graphene in the calculations, we used a rectangular periodic $(185 \times \sqrt{3})$ supercell that finally comprises two partials with alternating Burgers vectors (same edge components, opposite screw components). Furthermore, in the geometry optimization the atoms were either allowed to relax freely in all directions or were restricted to the x–y plane to reveal the pronounced effect of dislocation-induced buckling on the strain state of the membrane. From those configurations, the local atomistic strain, the disregistry between the two graphene layers, and the Burgers vector distributions for the edge and screw components of the partials were derived by evaluating the displacements of the carbon atoms.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 May; accepted 14 October 2013.
Published online 18 December 2013.

1. Mott, N. F. Dislocations and the theory of solids. *Nature* **171,** 234–237 (1953).
2. Orowan, E. Zur Kristallplastizität. III. Über den Mechanismus des Gleitvorganges. *Z. Phys.* **89,** 634–659 (1934).
3. Taylor, G. I. The mechanism of plastic deformation of crystals. Part II. Comparison with observations. *Proc. R. Soc. A* **145,** 388–404 (1934).
4. Cottrell, A. H. & Bilby, B. A. Dislocation theory of yielding and strain ageing of iron. *Proc. Phys. Soc. A* **62,** 49–62 (1949).
5. Frank, F. & Read, W. Multiplication processes for slow moving dislocations. *Phys. Rev.* **79,** 722–723 (1950).
6. Nabarro, F. R. N. Mathematical theory of stationary dislocations. *Adv. Phys.* **1,** 269–394 (1952).
7. Shockley, W. Dislocations and edge states in the diamond crystal structure. *Phys. Rev.* **91,** 228 (1953).
8. Smoluchowski, R. Dislocations in ionic crystals (structure, charge effects and interaction with impurities). *J. Phys. Colloq.* **27,** C3 (1966).
9. Hull, D. & Bacon, D. J. *Introduction to Dislocations* (Pergamon, 2001).
10. Waldmann, D. *et al.* Robust graphene membranes in a silicon carbide frame. *ACS Nano* **7,** 4441–4448 (2013).
11. Bao, W. *et al.* Stacking-dependent band gap and quantum transport in trilayer graphene. *Nature Phys.* **7,** 948–952 (2011).
12. Lui, C. H., Li, Z., Mak, K. F., Cappelluti, E. & Heinz, T. F. Observation of an electrically tunable band gap in trilayer graphene. *Nature Phys.* **7,** 944–947 (2011).
13. Wong, J.-H., Wu, B.-R. & Lin, M.-F. Strain effect on the electronic properties of single layer and bilayer graphene. *J. Phys. Chem. C* **116,** 8271–8277 (2012).
14. Vaezi, A., Liang, Y., Ngai, D. H., Yang, L. & Kim, E.-A. Topological edge states at a tilt boundary in gated multilayer graphene. *Phys. Rev. X* **3,** 021018 (2013).
15. Amelinckx, S. & Delavignette, P. Observation of dislocations in non-metallic layer structures. *Nature* **185,** 603–604 (1960).
16. Delavignette, P. & Amelinckx, S. Dislocation patterns in graphite. *J. Nucl. Mater.* **5,** 17–66 (1962).
17. Alden, J. S. *et al.* Strain solitons and topological defects in bilayer graphene. *Proc. Natl Acad. Sci. USA* **110,** 11256–11260 (2013).
18. Brown, L. *et al.* Twinning and twisting of tri- and bilayer graphene. *Nano Lett.* **12,** 1609–1615 (2012).
19. Hattendorf, S., Georgi, A., Liebmann, M. & Morgenstern, M. Networks of ABA and ABC stacked graphene on mica observed by scanning tunneling microscopy. *Surf. Sci.* **610,** 53–58 (2013).
20. Lin, J. *et al.* AC/AB stacking boundaries in bilayer graphene. *Nano Lett.* **13,** 3262–3268 (2013).
21. Shevitski, B. *et al.* Dark-field transmission electron microscopy and the Debye-Waller factor of graphene. *Phys. Rev. B* **87,** 045417 (2013).
22. Ping, J. & Fuhrer, M. S. Layer number and stacking sequence imaging of few-layer graphene by transmission electron microscopy. *Nano Lett.* **12,** 4635–4641 (2012).
23. Meyer, J. C. *et al.* The structure of suspended graphene sheets. *Nature* **446,** 60–63 (2007).
24. Warner, J. H. *et al.* Dislocation-driven deformations in graphene. *Science* **337,** 209–212 (2012).
25. Lehtinen, O., Kurasch, S., Krasheninnikov, A. V. & Kaiser, U. Atomic scale study of the life cycle of a dislocation in graphene from birth to annihilation. *Nature Commun.* **4,** 2098 (2013).
26. Yazyev, O. V. & Louie, S. G. Topological defects in graphene: dislocations and grain boundaries. *Phys. Rev. B* **81,** 195420 (2010).
27. Chen, S. & Chrzan, D. C. Continuum theory of dislocations and buckling in graphene. *Phys. Rev. B* **84,** 214103 (2011).
28. Eshelby, J. D. & Stroh, A. N. CXL. Dislocations in thin plates. *Phil. Mag.* **42,** 1401–1405 (1951).
29. Siems, R., Delavignette, P. & Amelinckx, S. The buckling of a thin plate due to the presence of an edge dislocation. *Phys. Status Solidi B* **2,** 421–438 (1962).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge use of experimental equipment of P. Schmuki. Furthermore, we thank J. Müller, E. Bitzek and A. Kohlmeyer for discussions. This research was supported by the Deutsche Forschungsgemeinschaft within the frameworks of the SFB 953 'Synthetic Carbon Allotropes' and the Cluster of Excellence EXC 315 'Engineering of Advanced Materials' at the Friedrich-Alexander-Universität Erlangen-Nürnberg.

Author Contributions B.B. and E.S. designed the experiments. D.W. and C.D. prepared the membranes based on the route recently developed and optimized by D.W., H.B.W. and B.B. B.B. and C.D. conducted the TEM experiments. B.B., E.S. and C.D. evaluated the experimental data. K.W. and B.M. performed the atomistic simulations. F.N. simulated the TEM dark-field images and rocking curves (using the atomistic configurations from K.W. and B.M.). Furthermore, F.N. determined the 2D strain and derived the Burgers vector distributions. B.B. and E.S. wrote the manuscript. All authors discussed the results and implications and commented on the manuscript at all stages.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.B. (benjamin.butz@ww.uni-erlangen.de) or E.S. (erdmann.spiecker@ww.uni-erlangen.de).

doi:10.1038/nature12945

Impacts of the north and tropical Atlantic Ocean on the Antarctic Peninsula and sea ice

Xichen Li[1], David M. Holland[1], Edwin P. Gerber[1] & Changhyun Yoo[1]

In recent decades, Antarctica has experienced pronounced climate changes. The Antarctic Peninsula exhibited the strongest warming[1,2] of any region on the planet, causing rapid changes in land ice[3,4]. Additionally, in contrast to the sea-ice decline over the Arctic, Antarctic sea ice has not declined, but has instead undergone a perplexing redistribution[5,6]. Antarctic climate is influenced by, among other factors, changes in radiative forcing[7] and remote Pacific climate variability[8,9], but none explains the observed Antarctic Peninsula warming or the sea-ice redistribution in austral winter. However, in the north and tropical Atlantic Ocean, the Atlantic Multidecadal Oscillation[10,11] (a leading mode of sea surface temperature variability) has been overlooked in this context. Here we show that sea surface warming related to the Atlantic Multidecadal Oscillation reduces the surface pressure in the Amundsen Sea and contributes to the observed dipole-like sea-ice redistribution between the Ross and Amundsen–Bellingshausen–Weddell seas and to the Antarctic Peninsula warming. Support for these findings comes from analysis of observational and reanalysis data, and independently from both comprehensive and idealized atmospheric model simulations. We suggest that the north and tropical Atlantic is important for projections of future climate change in Antarctica, and has the potential to affect the global thermohaline circulation[6] and sea-level change[3,12].

Recent multidecadal changes in Antarctic climate are well documented. Surface air temperature (SAT) in the Antarctic Peninsula[2,13], at the Faraday/Vernadsky station in particular, reveals a rapid warming trend of 5.6 K over 50 years in austral winter[1]. The accelerated land-ice melting around the Antarctic Peninsula and the Amundsen Sea[3,4] indicates enhanced warm air advection[3] and warm water transport to these regions[4]. Satellite observations show a dipole-like change in sea-ice concentration[5,6] (SIC), with increases in the Ross Sea[14] and decreases in the Amundsen–Bellingshausen–Weddell seas[6].

These regional-scale, multidecadal changes in Antarctica are strongly influenced by the atmospheric circulation[1,7]. During austral summer, Antarctic changes have been attributed to greenhouse gas increase[7] and stratospheric ozone loss[7,15,16], both of which project strongly onto the Southern Annular Mode[17,18]. In winter, however, the mechanisms driving Antarctic changes are less well understood. Previous studies have linked winter Antarctic climate variability to Pacific sea surface temperature (SST) variability, including the El Niño/Southern Oscillation[8,9] (ENSO) and central Pacific Ocean warming[19]. It is unclear, however, whether Pacific SST can explain Antarctic multidecadal climate changes. The ENSO does not exhibit significant multidecadal trends and therefore cannot account for recent Antarctic climate changes, and central Pacific warming appears to cool the Antarctic Peninsula and increase SIC over the Amundsen–Bellingshausen seas[19], in contrast to observed trends[1,2,5,6,14].

In comparison, the influence of the Atlantic Ocean has received less attention, although recent studies correlating Antarctic SAT with global SSTs have suggested potential links between Antarctic SAT and the south[2] and tropical[20] Atlantic. In this study, we single out north and tropical Atlantic SST as a key driver of recent Antarctic climate changes in austral winter, explaining both the Antarctic Peninsula warming and the SIC redistribution. Using observational data and numerical simulation, we demonstrate a teleconnection between the north and tropical Atlantic and Antarctica, and reveal the physical mechanism underlining this connectivity.

On decadal timescales, the Atlantic Multidecadal Oscillation[10,21] is a leading mode of global variability. Monthly mean north Atlantic SST variability since 1870 is dominated by a centennial warming (Fig. 1a) and a 60–70-year oscillation[10] (Fig. 1b). The Atlantic Multidecadal Oscillation is observed in SST reconstructions[10], palaeo-records[22] and millennial-scale climate simulations[21,23]. It has been associated with changes in the oceanic global thermohaline circulation[21,23], but may also be influenced by changes in atmospheric blocking[24] and anthropogenic forcing associated with the indirect aerosol effect[11,25]. The Atlantic Multidecadal Oscillation spatial pattern (Fig. 1e) exhibits two local maxima, one south of Greenland and one in the tropical Atlantic. To distinguish further the role of the tropical Atlantic, time series from this region alone are considered alongside north Atlantic SSTs in Fig. 1c. Tropical Atlantic SSTs are highly correlated ($R = 0.77$) with north Atlantic SSTs.

In the satellite era, from 1979 onwards, the Atlantic Multidecadal Oscillation manifests itself as an upward trend in north Atlantic SSTs, which, in combination with anthropogenic forcing, has led to a warming of more than half a degree (Fig. 1c). Although this positive trend gives the Atlantic Multidecadal Oscillation the potential to drive Antarctic climate changes, it complicates the assessment of the Atlantic Multidecadal Oscillation's impact in Antarctic observational records. To circumvent this limitation, we focus on detrended Atlantic SSTs (Fig. 1d), hereafter referred to as sub-decadal variability. Because these sub-decadal anomalies are slow relative to the timescales of the atmospheric wave propagation (Extended Data Fig. 1) but faster than the oceanic thermohaline circulation, we use them to identify the atmospheric fingerprint associated with warming in the north and tropical Atlantic. This is achieved by independent methods including regression, maximum covariance analysis (MCA) and numerical simulations.

Regression of austral winter (June, July and August) sea-level pressure (SLP) onto the north and tropical Atlantic SST time series (Fig. 2a, b) reveals a teleconnection between the Atlantic and Antarctica. Atlantic warming leads to dipolar changes in SLP, with increases south of Australia and decreases over Antarctica, in particular in the Amundsen Sea Low region. It therefore bears some resemblance to the positive phase of the Southern Annular Mode, with a pattern correlation of 0.84. The anomaly is barotropic, extending from the surface up to 200 hPa (Extended Data Fig. 2). A similar teleconnection is observed in all seasons except austral summer (Extended Data Fig. 3).

We independently identify this teleconnection between Atlantic SST and Southern Hemisphere SLP (Fig. 2c–e) through MCA. The first mode captures 55% of the squared covariance, and reveals a SLP spatial pattern (Fig. 2c) similar to the regression patterns (Fig. 2a, b). The SST pattern (Fig. 2d) is comparable to the Atlantic Multidecadal Oscillation, with broad warming across the north Atlantic and a bi-centre structure. The SST time series (Fig. 2e) is highly correlated ($R = 0.85$) with Atlantic sub-decadal variability (Fig. 1d). MCA thus simultaneously

[1]Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012, USA.

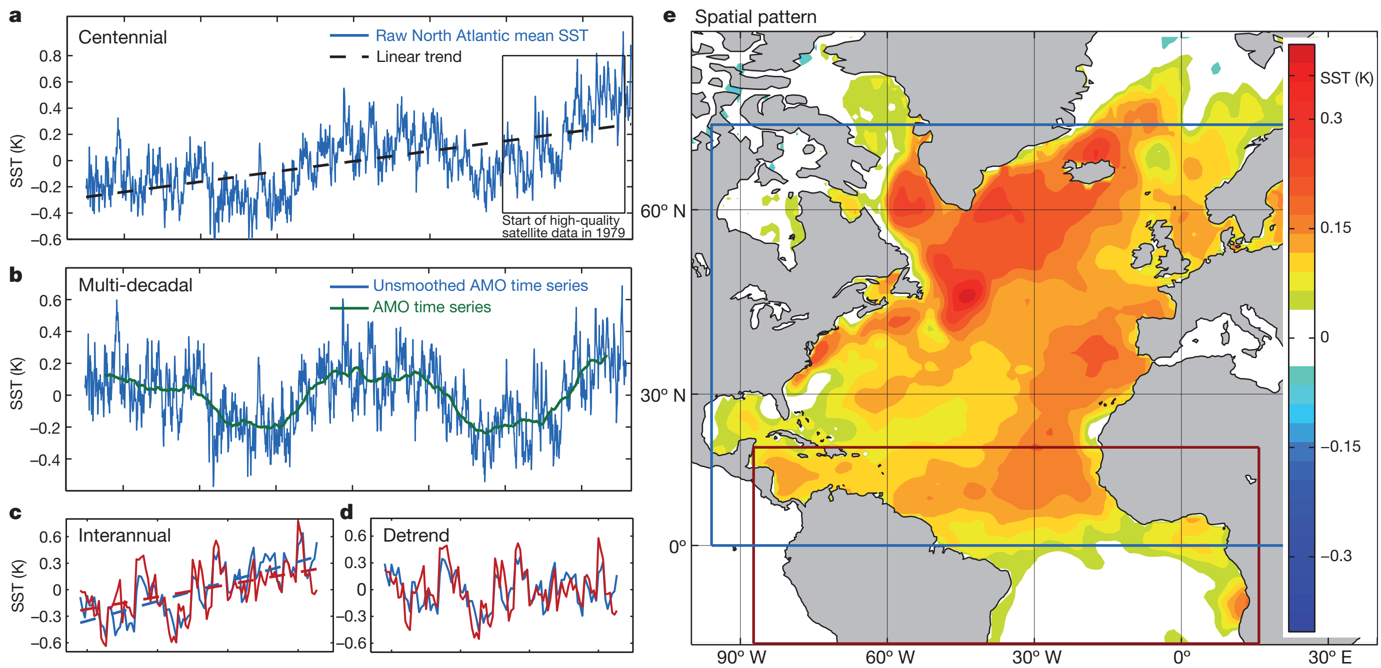


Figure 1 | Temporal variability of north and tropical Atlantic sea surface temperature anomalies at different timescales, and the spatial pattern of the Atlantic Multidecadal Oscillation. **a**, Area-weighted monthly mean SST in the north Atlantic (0° N–70° N) since 1870. The grey dashed line indicates the centennial trend. **b**, The blue curve shows mean SST after removing the linear trend signal and the green curve indicates the Atlantic Multidecadal Oscillation index, defined as a ten-year smoothed mean of the detrended time series. **c**, 1979–2012 (black box in **a**) austral winter monthly mean SST time series. The blue (or red) curve shows mean SST over the north (or tropical) Atlantic. The north (or tropical) Atlantic region is defined by the blue (or red) rectangle in **e**; the Pacific sector inside each box is excluded. The linear trends are indicated by the dashed lines in **c**, which are a superposition of the global warming trend in **a** and the ascending Atlantic Multidecadal Oscillation index in **b**. **d**, The detrended time series of **c** serve as indices of the north and tropical Atlantic sub-decadal variability and are used in regression. **e**, The spatial pattern of the Atlantic Multidecadal Oscillation, defined as the normalized regression of north Atlantic SST (1870–2012) against the Atlantic Multidecadal Oscillation index, exhibiting two warm centres, one south of Greenland and one in the tropics.

reproduces the Atlantic Multidecadal Oscillation pattern and associates it with the Antarctic SLP teleconnection.

We establish a causal link between Atlantic SSTs and the Antarctic SLP pattern with numerical simulations using the Community Atmosphere Model (CAM4), a state-of-the-art atmospheric model (see Methods for details). The simulated SLP response to north (Fig. 2f) and tropical (Fig. 2g) Atlantic warming is comparable to the SLP pattern obtained from regression and MCA. The model reproduces the amplification of the Amundsen Sea Low, albeit with an eastward shift.

The simulation results suggest that warming in the tropical Atlantic generates the bulk of the SLP response to the whole Atlantic Multidecadal Oscillation pattern (Fig. 2f, g). The SLP response to warming in the mid-latitude north Atlantic is comparatively weaker (Extended Data Fig. 4). The SLP response to the Atlantic Multidecadal Oscillation can be viewed as a linear combination of the responses to mid-latitude north and tropical Atlantic warming, with the latter playing the key part.

Geostrophic balance connects SLP changes to surface wind anomalies, which affect regional-scale changes in SIC and SAT[6,14,26]. Regression of SLP, SIC and SAT onto tropical Atlantic sub-decadal variability (Fig. 3a) reveals the impact of the atmospheric teleconnection on Antarctic SIC and SAT. In particular, the low-pressure anomaly in the Amundsen Sea induces a cyclonic (clockwise) circulation, heating the Antarctic Peninsula by warm-air advection[14] (red arrows in Fig. 3a). Changes in thermal advection and wind-stress forcing associated with surface wind anomalies drive a SIC dipole redistribution[6,26]. Local feedbacks between the sea ice and the ocean may considerably amplify the initial response triggered by the wind forcing[27,28].

To assess the statistical significance of the SIC and SAT patterns associated with Atlantic SSTs, we must also take into account the spatial structure of the response. The aggregated SIC increase across the Ross Sea, the SIC loss in the Amundsen–Bellingshausen–Weddell seas, and the SAT warming of the Antarctic Peninsula are statistically significant (see Methods and Extended Data Fig. 8).

Thus far we have established a climatic fingerprint of Atlantic warming on Antarctic climate (Fig. 3a) from sub-decadal variability. Because the timescales of the atmospheric circulation are fast relative to sub-decadal oceanic variability, the multidecadal response of Antarctic climate to Atlantic warming may also exhibit the same fingerprint. Multidecadal changes of SIC and SAT are shown in Fig. 3b. Reanalysis SAT is potentially biased over Antarctica[29], so we also include ground station SAT trends. SIC exhibits expansion around Antarctica, accompanied with a dipole redistribution between the Ross Sea and the Amundsen–Bellingshausen–Weddell seas[6,14] that is consistent with a deepening of the Amundsen Sea Low[14,18]. Both station observations and reanalysis SAT indicate a strong warming signal over the Antarctic Peninsula[1], although the cooling signal over Marie Byrd Land in the regression differs from observations[30], indicating that Atlantic SST changes do not explain the entire spatial pattern of west Antarctic warming. In east Antarctica, regression results (Fig. 3a) agree with station observations (Fig. 3b), both revealing a mild cooling trend. The reanalysis (Fig. 3b) indicates a warming tendency, but may be biased by surface energy balance errors[29].

A comparison of Fig. 3a and Fig. 3b reveals a strong resemblance between the sub-decadal fingerprint and the multidecadal changes in SIC and SAT. The SIC and SAT anomalies in Fig. 3a are associated with one standard deviation of the Atlantic SST variability, which is approximately 0.2 K. The net warming trend in the tropical Atlantic during the period 1979–2012, however, is approximately 0.5 K. Assuming a linear relationship between Atlantic SST and SIC/SAT, the amplitude of the fingerprint in Fig. 3a should be scaled by a factor of two, thus making it comparable to the multidecadal changes (Fig. 3b).

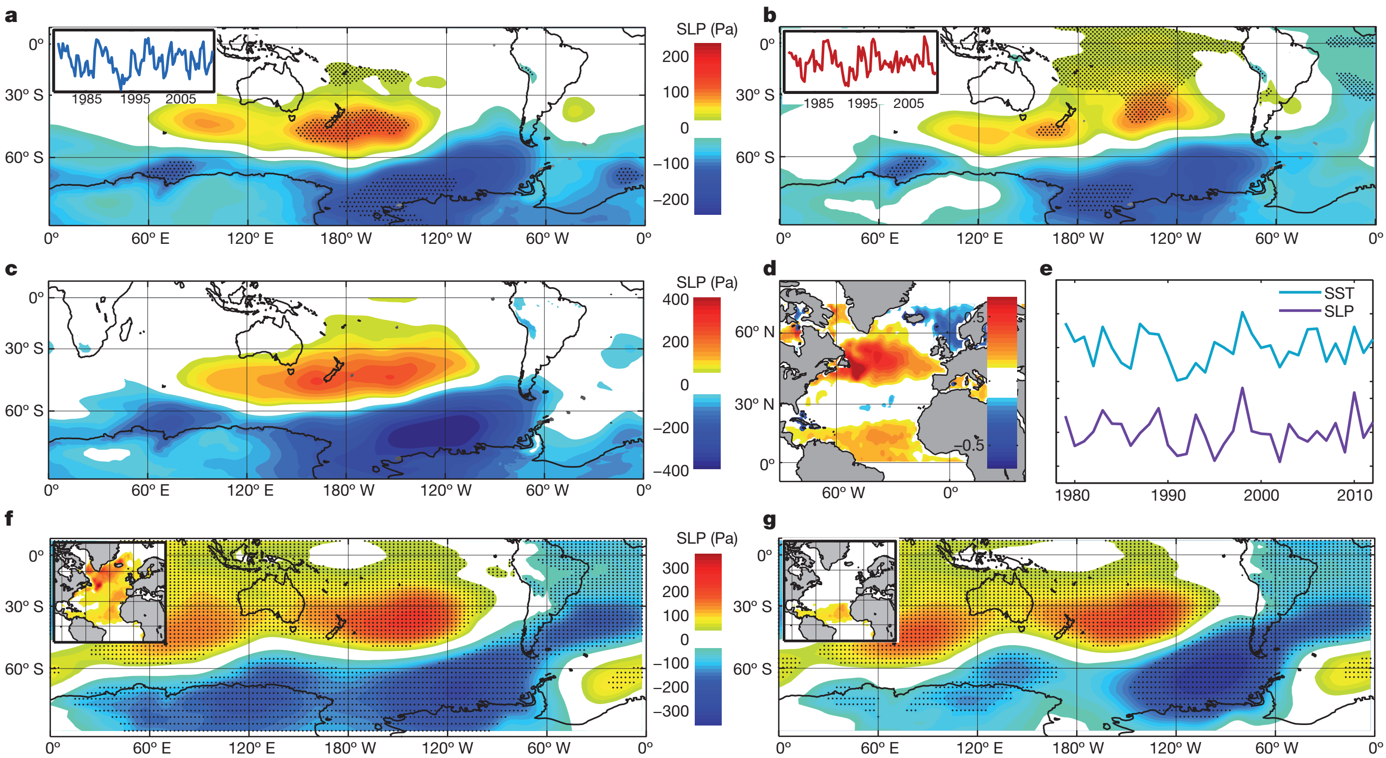


Figure 2 | North and tropical Atlantic variability projected onto austral winter Southern Hemisphere SLP by three independent analyses. **a**, **b**, The normalized regression of SLP against north Atlantic SST (**a**; the inset shows the time series) and tropical Atlantic SST (**b**). Areas of >95% significance (Student's *t*-test) are marked by black dots, restricted to the shaded regions. **c**–**e**, The first mode of MCA, including the spatial patterns of SLP (**c**) and SST (**d**), as well as their time series (**e**; unit free). **f**, **g**, Simulated SLP anomaly response to north Atlantic SST warming (**f**; the inset shows SST forcing) and tropical Atlantic SST warming (**g**; the inset shows SST forcing). The three analyses (**a**, **c** and **f**) show a coherent spatial pattern, suggesting a robust link between north Atlantic warming and Antarctic SLP and circulation anomalies. Evidence from simulation (**f**, **g**) implies an impact and causality coming from both the north and tropical Atlantic.

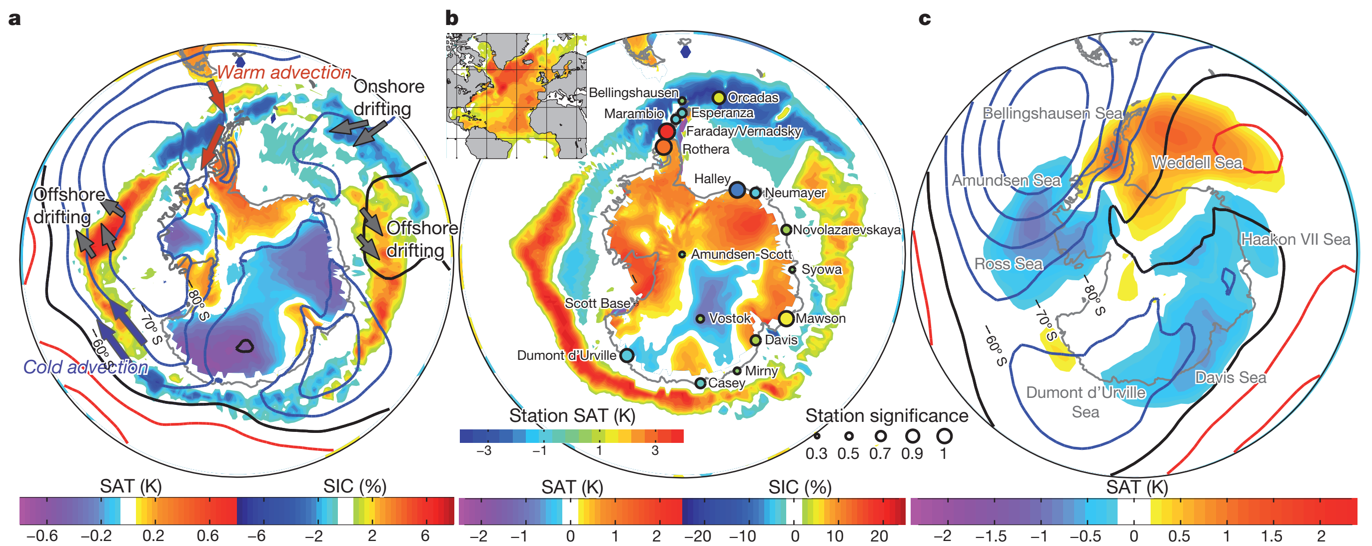


Figure 3 | The austral winter patterns of Antarctic SLP, SAT and SIC related to tropical Atlantic SST warming. **a**, SLP (red and blue contours indicate positive and negative anomalies respectively, 40 Pa interval), SAT (land-area colour) and SIC (ocean-area colour), individually regressed against the normalized tropical Atlantic SST. The regression on the north Atlantic SST shows similar signals, not shown. Atlantic SST-induced SAT/SIC patterns are consistent with SLP, via the mechanisms of thermal advection (red/blue arrows) and mechanical stress forcing (grey arrows). **b**, Epochal differences (1996–2012 minus 1979–1995) of SIC and SAT. SAT observations from 18 stations are also superimposed (circles, with colour showing the trends, and size corresponding to significance level). The Atlantic Multidecadal Oscillation spatial pattern is shown in the upper left. The similarity between sub-decadal variability (**a**) and multidecadal trends (**b**) implies a teleconnection between the Atlantic Multidecadal Oscillation and recent Antarctic climate change. The epochal difference of SLP is not shown in **b** because there exist uncertainties in the SLP reanalysis data. **c**, Simulated SLP (contours with 80 Pa interval) and SAT (colour) response to tropical Atlantic SST forcing (SIC is not included in the simulation) reveals causality in the above teleconnection.

The CAM4 simulations also suggest that the observed trend in SAT is a consequence of the circulation changes driven by tropical Atlantic warming. The SLP and SAT response to climatological tropical Atlantic SST forcing in CAM4 (Fig. 3c) nearly matches the regression results (Fig. 3a) and observed multidecadal trends (Fig. 3b). The amplitude of the CAM4 SAT patterns, however, should be interpreted with some caution. SAT in the polar region is affected by SIC changes, so the total response triggered by the atmospheric circulation change may depend on air–sea-ice–ocean interactions that are missing from our atmosphere-only model.

Remarkably, the SLP, SAT and SIC patterns obtained from sub-decadal regression, multidecadal observational trends and atmospheric model simulations all reveal a similar teleconnection between the north and tropical Atlantic and Antarctica. The regression and MCA results were derived using detrended, sub-decadal variability, and thus cannot independently inform us about multidecadal trends. However, given that the timescales of atmospheric variability are fast compared to sub-decadal climate variability, this sub-decadal teleconnection may imply the existence of a multidecadal link. The coherence between the multi-decadal trend (Fig. 3b) and detrended sub-decadal regression (Fig. 3a) significantly strengthens this argument, and the model simulations (Fig. 3c) establish causality, showing that Atlantic Multidecadal Oscillation SST forcing is a key driver of recent Antarctic climate change.

The mechanism by which tropical SSTs drive Antarctic atmospheric circulation depends critically on poleward propagating Rossby wave trains[8]. (Rossby waves are large-scale atmospheric wave structures that arise from variations in the effect of planetary rotation with latitude.) A similar physical process has been examined in previous studies[9,19], but focusing only on the tropical Pacific. In contrast, we simulate Rossby wave trains driven by tropical Atlantic SSTs (Fig. 2f, g and Extended Data Fig. 1g). CAM4 is a comprehensive atmospheric model that includes a wide array of physical processes, so it is difficult to isolate pure Rossby wave dynamics unambiguously. To focus on these dynamics alone, we performed idealized simulations with the 'dry dynamical core' of a Geophysical Fluid Dynamics Laboratory (GFDL) atmospheric model (see Methods for details). Using austral winter background conditions (Extended Data Fig. 1a–f), we demonstrate that convective heating over the tropical Atlantic generates Rossby wave trains that propagate around the globe within two weeks, ultimately focusing on and enhancing the Amundsen Sea Low. This wave pattern matches well with CAM4 simulations (Extended Data Fig. 1g), and further establishes that Rossby wave trains directly link the tropical Atlantic to Antarctica.

Although we have used subdecadal variability to identify the teleconnection from the north and tropical Atlantic to the Antarctic region, it is important to note that Pacific SST variability—the ENSO in particular—dominates the interannual variability of Antarctic climate[8,9,27]. The Atlantic, however, becomes a dominant driver of Antarctic climate on multidecadal timescales (see Methods and Extended Data Fig. 6).

In addition to the demonstrated impacts on SAT and SIC, this study implies broader impacts of north and tropical Atlantic warming, specifically on global sea-level change and the thermohaline circulation. The dramatic breakup of the Larsen A and B ice shelves and the present thinning of the C shelf have been attributed to atmospheric thermal advection[3], shown in this study to be driven, in part, by north and tropical Atlantic warming. These surface land-ice changes, in conjunction with basal melting processes caused by sub-ice-shelf intrusion of warm water, are contributing to an accelerated global sea-level rise[12]. Additionally, the SAT and SIC in the Weddell Sea, which we have shown to be sensitive to the Atlantic Multidecadal Oscillation, are critical to the formation of Antarctic bottom water[6], an important generator of the thermohaline circulation. The Atlantic Multidecadal Oscillation itself is considered to be the ocean surface response to changes in the thermohaline circulation[21,23]. This raises the possibility that the Atlantic Multidecadal Oscillation directly interacts with the thermohaline circulation in the Southern Ocean. Such an interaction may be important for understanding the climate system on multidecadal and even longer timescales.

METHODS SUMMARY

We used linear regression, MCA decomposition and two numerical models in this study. Statistical confidence levels are shown with linear regression indices, and MCA decomposition identifies the mode that maximizes the covariance between two different variables. The CAM4 response to Atlantic Multidecadal Oscillation SST forcing was obtained as the difference between a control run forced with 1980s boundary conditions and a perturbation run in which only SSTs in the north and tropical Atlantic were modified. The GFDL model shows the dynamical response to an idealized initial perturbation, mimicking the Atlantic Multidecadal Oscillation SST warming.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 August; accepted 10 December 2013.

1. Vaughan, D. G., Marshall, G. J., Connolley, W. M., King, J. C. & Mulvaney, R. Devil in the detail. *Science* **293,** 1777–1779 (2001).
2. Schneider, D. P., Deser, C. & Okumura, Y. An assessment and interpretation of the observed warming of West Antarctica in the austral spring. *Clim. Dyn.* **38,** 323–347 (2012).
3. Pritchard, H. *et al.* Antarctic ice-sheet loss driven by basal melting of ice shelves. *Nature* **484,** 502–505 (2012).
4. Joughin, I., Alley, R. B. & Holland, D. M. Ice-sheet response of oceanic forcing. *Science* **338,** 1172–1176 (2012).
5. Yuan, X. & Martinson, D. G. The Antarctic dipole and its predictability. *Geophys. Res. Lett.* **28,** 3609–3612 (2001).
6. Holland, P. R. & Kwok, R. Wind-driven trends in Antarctic sea-ice drift. *Nature Geosci.* **5,** 872–875 (2012).
7. Arblaster, J. M. & Meehl, G. A. Contributions of external forcings to southern annular mode trends. *J. Clim.* **19,** 2896–2905 (2006).
8. Karoly, D. J. Southern hemisphere circulation features associated with El Niño-Southern Oscillation events. *J. Clim.* **2,** 1239–1252 (1989).
9. Fogt, R. L., Bromwich, D. H. & Hines, K. M. Understanding the SAM influence on the South Pacific ENSO teleconnection. *Clim. Dyn.* **36,** 1555–1576 (2011).
10. Schlesinger, M. E. & Ramankutty, N. An oscillation in the global climate system of period 65–70 years. *Nature* **367,** 723–726 (1994).
11. Booth, B. B., Dunstone, N. J., Halloran, P. R., Andrews, T. & Bellouin, N. Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature* **484,** 228–232 (2012).
12. King, M. A. *et al.* Lower satellite-gravimetry estimates of Antarctic sea-level contribution. *Nature* **491,** 586–589 (2012).
13. O'Donnell, R., Lewis, N., McIntyre, S. & Condon, J. Improved methods for PCA-based reconstructions: case study using the Steig *et al.* (2009) Antarctic temperature reconstruction. *J. Clim.* **24,** 2099–2115 (2011).
14. Stammerjohn, S. E., Martinson, D. G., Smith, R. C., Yuan, X. & Rind, D. Trends in Antarctic annual sea ice retreat and advance and their relation to El Niño–Southern Oscillation and Southern Annular Mode variability. *J. Geophys. Res.* **113,** C03S90 (2008).
15. Son, S. W. *et al.* The impact of stratospheric ozone recovery on the Southern Hemisphere westerly jet. *Science* **320,** 1486–1489 (2008).
16. Turner, J. *et al.* Non-annular atmospheric circulation change induced by stratospheric ozone depletion and its role in the recent increase of Antarctic sea ice extent. *Geophys. Res. Lett.* **36,** L08502 (2009).
17. Thompson, D. *et al.* Signatures of the Antarctic ozone hole in Southern Hemisphere surface climate change. *Nature Geosci.* **4,** 741–749 (2011).
18. Marshall, G. J. Trends in the Southern Annular Mode from observations and reanalyses. *J. Clim.* **16,** 4134–4143 (2003).
19. Ding, Q., Steig, E. J., Battisti, D. S. & Küttel, M. Winter warming in West Antarctica caused by central tropical Pacific warming. *Nature Geosci.* **4,** 398–403 (2011).
20. Okumura, Y. M., Schneider, D., Deser, C. & Wilson, R. Decadal–interdecadal climate variability over antarctica and linkages to the tropics: analysis of ice core, instrumental, and tropical proxy data. *J. Clim.* **25,** 7421–7441 (2012).
21. Knight, J. R., Allan, R. J., Folland, C. K., Vellinga, M. & Mann, M. E. A signature of persistent natural thermohaline circulation cycles in observed climate. *Geophys. Res. Lett.* **32,** L20708 (2005).
22. Gray, S. T., Graumlich, L. J., Betancourt, J. L. & Pederson, G. T. A tree-ring based reconstruction of the Atlantic Multidecadal Oscillation since 1567 A.D. *Geophys. Res. Lett.* **31,** L12205 (2004).
23. Zhang, R. & Delworth, T. L. A new method for attributing climate variations over the Atlantic Hurricane Basin's main development region. *Geophys. Res. Lett.* **36,** L06701 (2009).
24. Häkkinen, S., Rhines, P. B. & Worthen, D. L. Atmospheric blocking and Atlantic multidecadal ocean variability. *Science* **334,** 655–659 (2011).
25. Ting, M., Kushnir, Y., Seager, R. & Li, C. Forced and internal twentieth-century SST Trends in the North Atlantic. *J. Clim.* **22,** 1469–1481 (2009).

26. Lefebvre, W. & Goosse, H. Influence of the Southern Annular Mode on the sea ice-ocean system: the role of the thermal and mechanical forcing. *Ocean Sci. Discuss.* **2,** 299–329 (2005).
27. Stammerjohn, S., Massom, R., Rind, D. & Martinson, D. Regions of rapid sea ice change: An inter-hemispheric seasonal comparison. *Geophys. Res. Lett.* **39,** L06501 (2012).
28. Goosse, H. & Zunz, V. Decadal trends in the Antarctic sea ice extent ultimately controlled by ice-ocean feedback. *Cryosphere Discuss.* **7,** 4585–4632 (2013).
29. Bracegirdle, T. J. & Marshall, G. J. The reliability of Antarctic tropospheric pressure and temperature in the latest global reanalyses. *J. Clim.* **25,** 7138–7146 (2012).
30. Bromwich, D. H. *et al.* Central West Antarctica among the most rapidly warming regions on Earth. *Nature Geosci.* **6,** 139–145 (2013).

Acknowledgements X.L., D.M.H. and C.Y. were supported by the NSF Office of Polar Programs (grant number ANT-0732869), the NASA Polar Programs (grant number NNX12AB69G), and New York University Abu Dhabi (grant number G1204). E.P.G. was supported by the NSF Office of Atmospheric and Geospace Sciences (grant number AGS-1264195). The HadISST SST and SIC data was provided by the British Met Office, Hadley Centre. The Antarctic weather station data was made available by the British Antarctic Survey. The MERRA atmospheric reanalysis data was provided by the Global Modeling and Assimilation Office (GMAO) at NASA Goddard Space Flight Center (GSFC) through the NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC) online archive (http://disc.sci.gsfc.nasa.gov/mdisc/data-holdings/merra/merra_products_nonjs.shtml). The ERA-Interim atmospheric reanalysis was provided by the ECMWF. The comprehensive atmospheric model (CAM4) was made available by the National Center for Atmospheric Research (NCAR), supported by the National Science Foundation (NSF) and the Office of Science (BER) of the US Department of Energy (DOE). The idealized atmospheric model (the GFDL dry dynamical core) was developed by the National Oceanic and Atmospheric Administration (NOAA) at the GFDL. Computing resources were provided by the National Energy Research Scientific Computing Center (NERSC) and High Performance Computing (HPC) at New York University (NYU).

Author Contributions X.L., D.M.H. and E.P.G. designed the experiments; X.L. performed the data analysis and CAM4 numerical simulations, and prepared all figures; C.Y. ran the initial value calculations; C.Y. and X.L. created Extended Data Fig. 1 and all authors wrote and reviewed the main manuscript text.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.L. (xichen@cims.nyu.edu).

doi:10.1038/nature12901

Mycorrhiza-mediated competition between plants and decomposers drives soil carbon storage

Colin Averill[1], Benjamin L. Turner[2] & Adrien C. Finzi[3]

Soil contains more carbon than the atmosphere and vegetation combined[1]. Understanding the mechanisms controlling the accumulation and stability of soil carbon is critical to predicting the Earth's future climate[2,3]. Recent studies suggest that decomposition of soil organic matter is often limited by nitrogen availability to microbes[4–6] and that plants, via their fungal symbionts, compete directly with free-living decomposers for nitrogen[6,7]. Ectomycorrhizal and ericoid mycorrhizal (EEM) fungi produce nitrogen-degrading enzymes, allowing them greater access to organic nitrogen sources than arbuscular mycorrhizal (AM) fungi[8–10]. This leads to the theoretical prediction that soil carbon storage is greater in ecosystems dominated by EEM fungi than in those dominated by AM fungi[11]. Using global data sets, we show that soil in ecosystems dominated by EEM-associated plants contains 70% more carbon per unit nitrogen than soil in ecosystems dominated by AM-associated plants. The effect of mycorrhizal type on soil carbon is independent of, and of far larger consequence than, the effects of net primary production, temperature, precipitation and soil clay content. Hence the effect of mycorrhizal type on soil carbon content holds at the global scale. This finding links the functional traits of mycorrhizal fungi to carbon storage at ecosystem-to-global scales, suggesting that plant–decomposer competition for nutrients exerts a fundamental control over the terrestrial carbon cycle.

Nitrogen (N) availability influences biosphere–atmosphere exchanges of carbon (C) by limiting C inputs to the soil from net primary production[12] (NPP), and C outputs associated with the activity of decomposer microbes[4]. Most plant species on the Earth associate with symbiotic mycorrhizal fungi to acquire nutrients from soil. EEM fungi produce a wide range of enzymes that release N from soil organic matter[13], whereas AM fungi lack these enzyme systems[10,14]. Accordingly, EEM-associated plants (EEM plants) acquire substantially more organic N from the soil than do AM-associated plants (AM plants)[9,10,13], and also compete directly for organic N with other free-living decomposer microbes in the soil. A recent theoretical model suggests that uptake of organic N by EEM plants slows the rate of decomposition and increases soil C storage by exacerbating the nitrogen limitation of free-living decomposer activity and their production of enzymes that degrade soil organic matter[11]. There is as yet little support for this contention.

We used a mixed effects model to test the hypothesis that ecosystems dominated by EEM fungi (EEM ecosystems) store significantly more soil C than do ecosystems dominated by AM fungi (AM ecosystems) after accounting for variation in soil N and other drivers of soil C storage. We assembled a global data set containing soil C, N and clay content to a depth of one metre, as well as site-specific vegetation descriptions to determine biome and mycorrhizal type (Table 1). We then used global data products to assign mean annual temperature (MAT), mean annual precipitation (MAP)[15], and NPP[16] to determine whether the effect of mycorrhizal type on soil C storage was statistically significant after accounting for variations in biome type and biophysical properties assumed to control decomposition in ecosystem and Earth system models[17] (see Methods Summary). The statistical analysis included soil clay content because such secondary minerals have the potential to sorb and stabilize soil organic carbon[18].

We found that EEM ecosystems store 1.7 times more C per unit of soil N than do AM ecosystems (Fig. 1). The most parsimonious, corrected Akaike Information Criterion (AICc)-selected model (mycorrhizal type $\times$ N interaction, $P < 0.0001$, AICc-selected model $R^2_{\mathrm{LR}} = 0.91$; Fig. 1 and Methods) supported the removal of climate variables, NPP and clay content, which were weakly correlated with soil C content, though the biome type remained in the model as a random effect (Fig. 2). The mycorrhizal $\times$ N interaction remained significant even when all predictors were included in the model. This result shows that mycorrhizal status exerts a far larger control over soil C content than do climate variables, NPP or clay content. Weak relationships between NPP, climate and soil C storage at the global scale have also been reported elsewhere[19].

We conducted a sensitivity analysis of the model to ensure the findings were robust to the exclusion of biomes that contained only a single mycorrhizal type, surface organic horizons (which represent a large fraction of surface soil C content) in cold climates and EEM data points whose soil N content was greater than the largest observation found in AM ecosystems (see Extended Data Figs 1–5 and Extended Data Tables 1–5). In all cases a mycorrhizal effect was retained. This suggests that current model formulations of the terrestrial C cycle—that is, NPP-driven accumulations of C in soil pools that turn over based on

Table 1 | Biome data

Biome	n	AM	EEM	C stock (kg C m^{-2})	N stock (kg N m^{-2})	MAT (°C)	MAP (mm)	NPP (kg C m^{-2} yr^{-1})	Clay (%)
Boreal forest	12	0	12	61.4 (11.7)	2.9 (0.5)	−2.5 (1.3)	497 (76)	319 (27)	5.8 (0.8)
Temperate forest	99	41	58	24.5 (2.6)	1.1 (0.1)	8.6 (0.4)	1,544 (133)	633 (28)	13.8 (1.1)
Tropical forest	104	83	21	11.7 (0.6)	1.1 (0.1)	24.7 (0.3)	2,697 (58)	956 (23)	47.9 (1.9)
Grassland	12	12	0	14.5 (2.2)	1.4 (0.3)	10.8 (1.7)	857 (145)	576 (99)	20.8 (3.1)

n, Number of observations; AM, number of *n* that are AM; EEM, number of *n* that are EEM; C stock, mean soil C content; N stock, mean soil N content; MAP (mm), MAT (°C), Clay, soil clay content. All values are means within a biome type, with the associated standard error given in parentheses.

[1]Department of Integrative Biology, Graduate Program in Ecology, Evolution and Behavior, University of Texas at Austin, Austin, Texas 78712, USA. [2]Smithsonian Tropical Research Institute, Apartado 0843-03092, Balboa, Ancon, Republic of Panama. [3]Department of Biology, Boston University, Boston, Massachusetts 02215, USA.

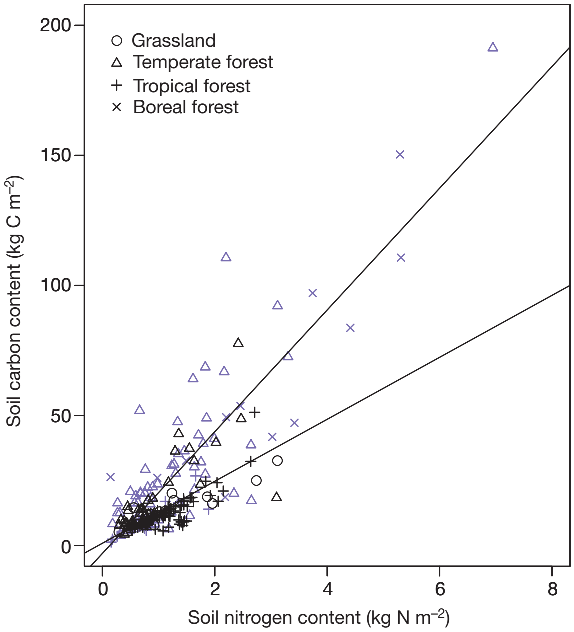


Figure 1 | The relationship between soil carbon and nitrogen content to a depth of one metre in AM and EEM ecosystems. The difference between the slopes is significant at the $P < 0.0001$ level, based on the best AICc-selected full model output, after starting with all predictors ($n = 227$). EEM systems store 1.7 times more C per unit N than do AM systems. Symbol shape reflects biome categorization. Symbol colour reflects mycorrhizal type, with purple symbols for EEM observations and black symbols for AM observations. Plotted lines represent univariate regression lines of the respective subsets of the data and are included for visualization purposes only.

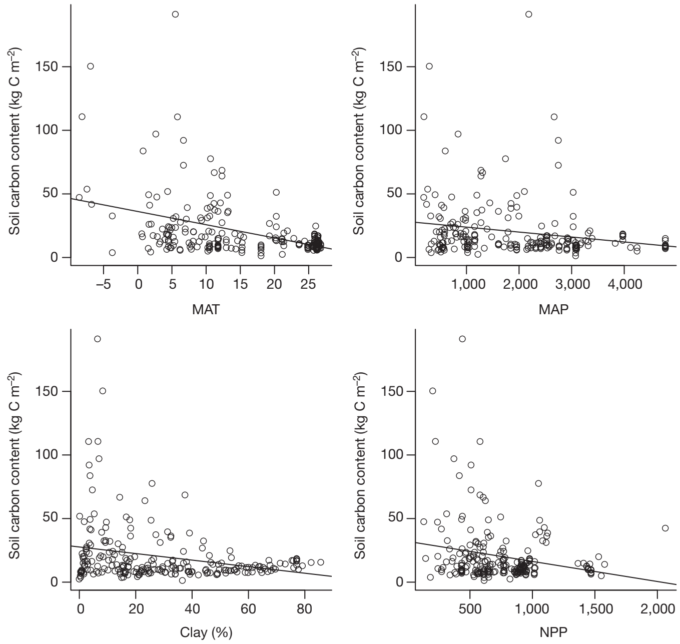


Figure 2 | The relationships between soil carbon content to a depth of one metre and MAT, MAP, clay content and NPP. Univariate regressions show that MAT in degrees celcius (**a**), MAP in mm precipitation per year (**b**), depth weighted clay content (**c**) and NPP in kg C m^{-2} per year (**d**) are significantly correlated with soil C storage ($R^2 = 0.18$, 0.04, 0.06 and 0.04, $P < 0.0001$, 0.0022, <0.0001 and 0.0009, respectively; $n = 227$). Regression lines represent univariate relationships rather than the output of the full model and are for visualization purposes only. None of these predictors were significant in the full model and were removed from the model after AICc selection.

Arrhenius temperature kinetics and a soil moisture multiplier[20]—lack a major driver of the decomposition process, namely, mycorrhizal type.

The results reported here support the recent theoretical contention that competition for organic N between EEM fungi and free-living microbes increases soil C storage[11], and we show that this effect holds from tropical to high-latitude ecosystems (Fig. 1). Competition-induced declines in decomposition rate in EEM ecosystems are further supported by natural abundance and ^{15}N-labelling studies that show that EEM plants acquire more organic N from the soil than do AM plants[13,21], and that experimental exclusion of EEM fungi increases the rate of organic matter decomposition[7] by increasing the biomass of free-living microbes and the activity of their C-degrading enzymes[6]. In contrast, the exclusion of AM fungi from the soil reduces the rate of decomposition by reducing the substrate supply to free-living decomposers[22]. Greater soil C storage in EEM ecosystems than in AM ecosystems at the large spatial scale reported here demonstrates that fine-scale mechanistic studies on the functionality of mycorrhizal symbioses—including N uptake preferences[13,21], partitioning of plant-C belowground[23], productivity[24] and decomposition[7,22]—can be scaled up to predict the consequences of AM versus EEM symbioses at the ecosystem-to-global scale.

It is possible that mycorrhizal effects on soil C pools may be confounded by differences in litter chemistry between EEM and AM plants. Compared to AM plants, litter from EEM plants can contain wider C:N ratios and greater concentrations of lignin and polyphenolic compounds, all of which are negatively correlated with short-term rates of litter decay[25]. However, lignin and polyphenols represent only a fraction of the soil C pool and compound-specific ^{13}C labelling studies show that these compounds decompose as fast or faster than 'labile' soil C compounds such as proteins and polysaccharides[26]. Moreover, recent theoretical[27] and empirical[28] work suggests that more recalcitrant (that is, more difficult to decompose) or higher C:N plant inputs may lead to less, rather than more, soil C storage than labile inputs because of lower microbial C-use efficiency (that is, the fraction of C assimilated that is allocated to growth rather than respiration). Therefore, we discount the possibility of a direct effect of litter chemistry on the observed variation in soil C storage among mycorrhizal types, although we cannot discount a potential indirect effect of litter chemistry owing to variations in microbial C-use efficiency.

This analysis shows that mycorrhizal functional traits are as important a control over decomposition and soil C storage as are soil chemical properties and the physical protection of organic matter[26]. More broadly, we demonstrate that the identity and functional traits of soil microorganisms exert a fundamental control over the terrestrial C cycle. This implies that global changes (for example, atmospheric N deposition, climate warming) that alter competitive interactions for N between EEM fungi and free-living microbial decomposers will affect soil C storage at regional to global scales.

METHODS SUMMARY

We collected data on soil C, N and percentage clay to a depth of one metre in soil profiles spanning tropical, temperate and boreal forests and grasslands. Data were obtained from a variety of sources, including direct observations and both published and unpublished data (see Acknowledgements and Methods). MAT, MAP and NPP were assigned on the basis of latitude and longitude using global data products. Data are summarized by biome in Table 1. Data were analysed in a mixed effects framework using the *lme* function implemented in the *nlme* package for R statistical software[29]. We tested for a main effect of mycorrhizal status on soil C as well as an interaction between mycorrhizal type and soil N with biome coded as a random effect and all other variables coded as fixed effects. Model selection was performed using AICc selection criteria to prevent over-fitting the model. Results discussed in the text are based on the full model output based on the best AICc-selected model starting with all predictors. Reported correlation factor R^2 values are based on the log ratio R^2. The mycorrhizal effect size reported in the text is determined by comparing the parameter estimate of the interaction between mycorrhizal type and soil N to the parameter estimate of the main effect of soil N, based on model outputs from the best AICc-selected full model.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 December 2012; accepted 25 November 2013.

Published online 8 January 2014.

1. Tarnocai, C. *et al.* Soil organic carbon pools in the northern circumpolar permafrost region. *Glob. Biogeochem. Cycles* **23,** GB2023 (2009).
2. Jenkinson, D. S., Adams, D. E. & Wild, A. Model estimates of CO_2 emissions from soil in response to global warming. *Nature* **351,** 304–306 (1991).
3. Knorr, W., Prentice, I. C., House, J. I. & Holland, E. A. Long-term sensitivity of soil carbon turnover to warming. *Nature* **433,** 298–301 (2005).
4. Schimel, J. P. & Weintraub, M. N. The implications of exoenzyme activity on microbial carbon and nitrogen limitation in soil: a theoretical model. *Soil Biol. Biochem.* **35,** 549–563 (2003).
5. Allison, S. D., Gartner, T. B., Mack, M. C., McGuire, K. & Treseder, K. Nitrogen alters carbon dynamics during early succession in boreal forest. *Soil Biol. Biochem.* **42,** 1157–1164 (2010).
6. Lindahl, B. D., de Boer, W. & Finlay, R. D. Disruption of root carbon transport into forest humus stimulates fungal opportunists at the expense of mycorrhizal fungi. *ISME J.* **4,** 872–881 (2010).
7. Gadgil, R. & Gadgil, P. Mycorrhiza and litter decomposition. *Science* **233,** 133 (1971).
8. Näsholm, T. *et al.* Boreal forest plants take up organic nitrogen. *Nature* **392,** 914–916 (1998).
9. Hodge, A., Campbell, C. D. & Fitter, A. H. An arbuscular mycorrhizal fungus accelerates decomposition and acquires nitrogen directly from organic material. *Nature* **413,** 297–299 (2001).
10. Read, D. J. & Perez-Moreno, J. Mycorrhizas and nutrient cycling in ecosystems—a journey towards relevance? *New Phytol.* **157,** 475–492 (2003).
11. Orwin, K. H., Kirschbaum, M. U. F., St John, M. G. & Dickie, I. A. Organic nutrient uptake by mycorrhizal fungi enhances ecosystem carbon storage: a model-based assessment. *Ecol. Lett.* **14,** 493–502 (2011).
12. LeBauer, D. S. & Treseder, K. K. Nitrogen limitation of net primary productivity in terrestrial ecosystems is globally distributed. *Ecology* **89,** 371–379 (2008).
13. Averill, C. & Finzi, A. Increasing plant use of organic nitrogen with elevation is reflected in nitrogen uptake rates and ecosystem $\delta^{15}N$. *Ecology* **92,** 883–891 (2011).
14. Lindahl, B. D., Finlay, R. D. & Cairney, J. W. G. Enzymatic activities of mycelia in mycorrhizal fungal communities. *The Fungal Community: its Organization and Role in the Ecosystem* 3rd edn, 331–348 (CRC Press, 2005).
15. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. E. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25,** 1965–1978 (2005).
16. Zhao, M. & Running, S. W. Drought-induced reduction in global terrestrial net primary production from 2000 through 2009. *Science* **329,** 940–943 (2010).
17. Randerson, J. T. *et al.* Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models. *Glob. Change Biol.* **15,** 2462–2484 (2009).
18. Torn, M. S., Trumbore, S. E., Chadwick, O. A., Vitousek, P. M. & Hendricks, D. M. Mineral control of soil organic carbon storage and turnover. *Nature* **389,** 170–173 (1997).
19. Cebrian, J. & Duarte, C. M. Plant growth-rate dependence of detrital carbon storage in ecosystems. *Science* **268,** 1606–1608 (1995).
20. Davidson, E. A. & Janssens, I. A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **440,** 165–173 (2006).
21. Gallet-Budynek, A. *et al.* Intact amino acid uptake by northern hardwood and conifer trees. *Oecologia* **160,** 129–138 (2009).
22. Cheng, L. *et al.* Arbuscular mycorrhizal fungi increase organic carbon decomposition under elevated CO_2. *Science* **337,** 1084–1087 (2012).
23. Rygiewicz, P. T. & Andersen, C. P. Mycorrhizae alter quality and quantity of carbon allocated below ground. *Nature* **369,** 58–60 (1994).
24. van der Heijden, M. G. A. *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396,** 69–72 (1998).
25. Talbot, J. M. & Finzi, A. C. Differential effects of sugar maple, red oak, and hemlock tannins on carbon and nitrogen cycling in temperate forest soils. *Oecologia* **155,** 583–592 (2008).
26. Schmidt, M. W. I. *et al.* Persistence of soil organic matter as an ecosystem property. *Nature* **478,** 49–56 (2011).
27. Cotrufo, M. F., Wallenstein, M. D., Boot, C. M., Denef, K. & Paul, E. The Microbial Efficiency-Matrix Stabilization (MEMS) framework integrates plant litter decomposition with soil organic matter stabilization: do labile plant inputs form stable soil organic matter? *Glob. Change Biol.* **19,** 988–995 (2013).
28. Frey, S. D., Lee, J., Melillo, J. M. & Six, J. The temperature response of soil microbial efficiency and its feedback to climate. *Nature Clim. Change* **3,** 395–398 (2013).
29. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & the R Development Core Team *nlme: Linear and Nonlinear Mixed Effects Models* R package version 3. 1–109, http://cran.r-project.org/web/packages/nlme/nlme.pdf (2013).

Acknowledgements We thank L. Nave and the International Soil Carbon Network for access to their database. C. Hawkes provided feedback during data collection and initial analyses of C storage. C. Iversen, J. Powers and M. Vadeboncouer provided unpublished data that contributed to this analysis. D. Jacquier provided the Australian soil database and E. Carlston helped to extract data from the Australian soil database. C. Shaw provided the Siltanen soil carbon database and the Forest Ecosystem Carbon Database of Canadian soils. T. Baisden provided scans of the California Soil-Vegetation Survey. E. Brzostek, N. Fowler, P. Groffman, E. Hobbie, B. Schlesinger and B. Waring provided feedback on earlier versions of this manuscript. The Center for Tropical Forest Science (CTFS) and Smithsonian Institution Geo-observatories (SIGEO) provided funding for the collection and analysis of soil profile data at large forest dynamics plots, and we thank the many collaborators, field assistants and laboratory technicians who assisted in the collection and analysis of soil profile data. This work benefited from extensive data contributions to the International Soil Carbon Network from both the USDA Natural Resources Conservation Service, National Cooperative Soil Survey, and the US Geological Survey. C.A. was supported by a fellowship from the University of Texas at Austin and by the National Science Foundation Graduate Research Fellowship Program (grant DGE-1110007). A.C.F. was supported by NSF grant number DEB 07-43564 and DOE grants 10-DOE-1053 and DE-SC0006916. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Author Contributions C.A. and B.L.T. collected the data. C.A. performed all statistical analyses. C.A. and A.C.F. conceptualized the work and wrote the manuscript. All authors contributed to revisions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.A. (colin.averill@utexas.edu).

METHODS

Data collection. We accessed the International Soil Carbon database in February 2012 (http://www.fluxdata.org/NSCN/SitePages/ISCN.aspx), which contained a substantial amount of data from the United States Department of Agriculture and the National Resources Conservation Service (http://ssldata.nrcs.usda.gov). Plant species descriptions were supplemented with the National Soil Information System (NASIS) database (available from the United States Department of Agriculture, National Resources Conservation Service on request). We supplemented this data with unpublished data on 103 soil profiles from temperate and tropical forests. We then further supplemented this database by accessing the Canadian Forest Ecosystem Carbon Database and the Siltanen Database provided by C. Shaw, the Australian National Soil Database provided by D. Jacquier and a subset of the California Soil-Vegetation Survey provided by T. Baisden. We further supplemented the tropical sites by performing a Google Scholar search for articles containing all of the words "bulk density" + "clay" + "nitrogen" + "meter", with the exact phrase: "soil carbon storage"; and with at least one of the words "tropics" and "tropical". We also put out calls for these data on the ESA Biogeosciences Listserv and the National Soil Carbon Network listserv. J. Powers and M. Vadeboncouer provided unpublished data for this analysis. C. Iversen provided unpublished data from sites described in ref. 30, with permission.

Calculating soil C and N content and clay concentration. Soil C and N content were estimated as the sum of bulk density weighted soil C or N values to a depth of one metre or bedrock. Soil C and N content to a depth of one metre included organic horizons. Percentage clay to a depth of one metre was calculated as the depth-weighted average perecentage clay concentration across all depth increments. Organic horizons were not included in the clay calculation. For our unpublished data set, soil C and N concentrations were determined by combustion and gas chromatography with thermal conductivity detection on a Thermo Flash 1112 Analyzer (CE Elantech), bulk density was determined by the excavation method[31], and particle size distribution (including clay content) was determined by the pipette method following pretreatment to remove soluble salts, organic matter and iron oxides[32].

Assigning NPP, MAT and MAP. NPP, MAT and MAP were assigned using global data products and the latitude and longitude of each site. NPP was determined from a ten-year average Moderate Resolution Imaging Spectroradiometer (MODIS) NPP product, MOD17A3 (ref. 16). MAT and MAP were taken from the WorldClim global climate data product[15].

Mycorrhizal classification. Mycorrhizal type was assigned based on a site-specific vegetation observation of dominant species. We excluded observations that described a mixed mycorrhizal composition when no relative abundance data was available, although for forest biomes we ignored understory plants. When relative abundance data were available we required at least 70% of basal area of trees exceeding 10 cm in diameter at breast height to be one mycorrhizal type or the other. Some vegetation descriptions merely said "grassland" or "plains", which we classified as AM. Seven observations had a vegetation description of "Sierran mixed coniferous forest", which we classified as EEM. We note that forest classifications did not always include information on understory species. Biomass of understory plants is quite small by contrast to canopy trees, so it is unlikely that the understory plants had an important effect on patterns of soil C storage.

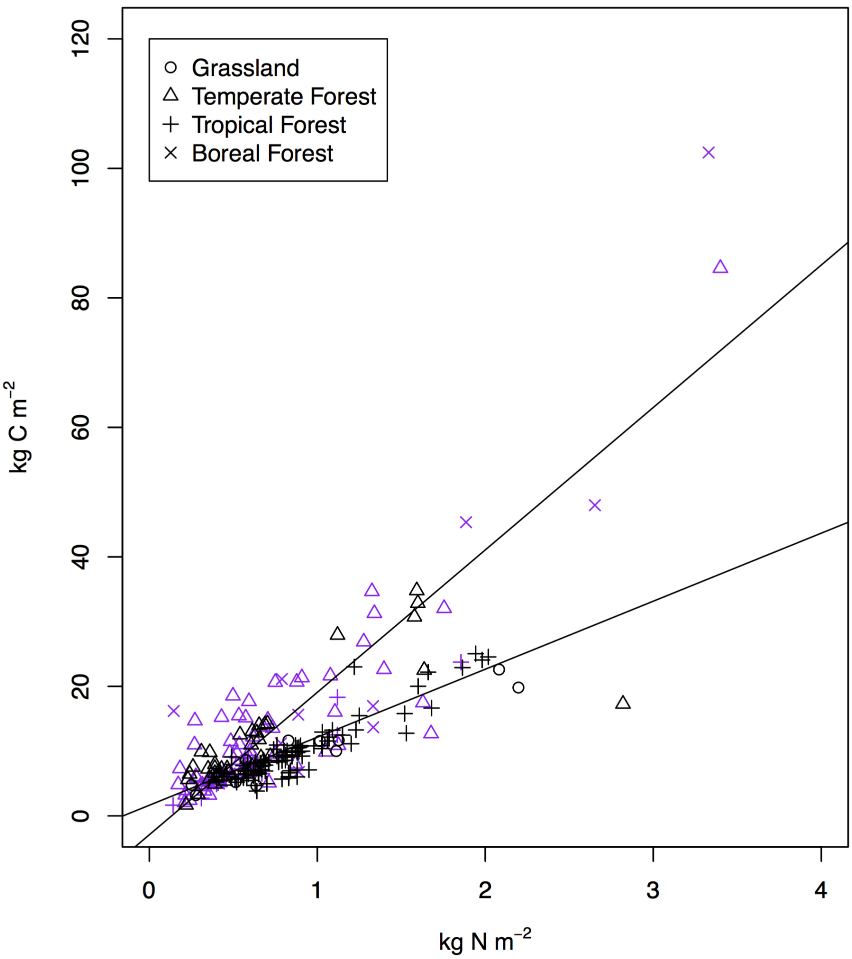
Biome classification. Biomes were assigned using the Whittaker Biome Diagram[33] and the MAT and MAP observations generated for each site from the WorldClim data product[15]. There were 32 instances in which Whittaker biome classifications were reassigned. Each of these soil observations was from a data set which had a description of vegetation and its United States Environmental Protection Agency ecoregion. If the Whittaker biome classification was not consistent with the vegetation description and EPA ecoregion, the biome was assigned to best match the vegetation description. For example, if the Whittaker MAT and MAP classification placed an observation into the temperate forest biome, but the vegetation description listed a "grassland" and the EPA ecoregion was "Great Plains," then we classified the observation as the grassland biome. Finally, 12 observations from temperate rain forest fell far outside the Whittaker biome diagram as they had exceptionally high MAP values for a temperate forest (>3,000 mm); however, we included them within the temperate forest category.

Statistical approach. We sought to model C storage as a function of mycorrhizal status and soil N while simultaneously accounting for variations in MAT, MAP, NPP and soil clay content. The data were analysed using a mixed effects framework, with MAT, MAP, NPP, clay, soil N, mycorrhizal status, and the interaction between N and mycorrhizal status as fixed effects and biome as a random effect because the number of AM and EM observations was not evenly distributed among biomes. Model selection was performed using corrected AIC (AICc) criteria, using the *AICcmodavg* package in R[34]. We required a minimum of a one-point AICc improvement to justify removing a term from the model. Final linear models did not have normally distributed residuals and were strongly heteroscedastic. The heteroscedasticity in the model probably arises from a sampling error that is a constant percentage of total observed soil C and N, rather than a constant value (that is, $\pm 10\%$ rather than ± 10 kg). We therefore fitted models by percentage least squares by weighting each observation by the inverse of the dependent variable (soil C stock) (as in ref. 35) and implemented using the weights function in *lme*[29]. Normality and homoscedasticity were inspected using plots of the normalized residuals. Because the R^2 metric has different properties in linear mixed effects models and in standard linear models, we report an R^2 statistic based on the likelihood ratio of the model (R^2_{LR}) that has properties similar to those of the R^2 implemented in linear models, as presented in ref. 36 and implemented in R using the *lmmfit* package[37]. The mycorrhizal effect size reported in the main text and Methods was determined by comparing the parameter estimate of the interaction between mycorrhizal type and soil N to the parameter estimate of the main effect of soil N, based on model outputs from the best AICc-selected full model.

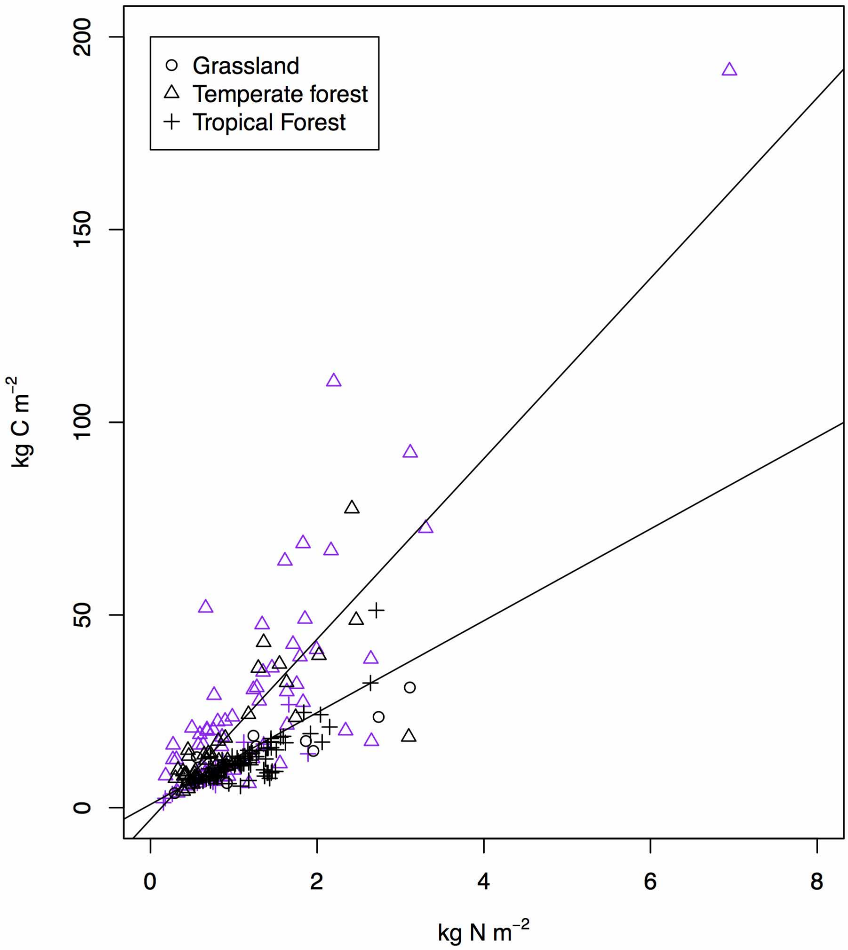
Multicollinearity was assessed using variance inflation factors. These were calculated using the *vif* function in the *car* package in R[38]. The factors were calculated from the linear ordinary least-squares regression of all fixed effects without interactions. Collinearity was determined to be not a problem because all variance inflation factor values were less than ten for all independent variables[39].

We further assessed the validity of the mixed effects framework by conducting a Monte Carlo simulation of the data set and analysis. Soil C distributions within biome and in most data sets are skewed distributions. To generate simulated data we determined the skew parameters of the soil C distribution (location, shape and scale), and then simulated a new biome-specific soil C distribution using the *rsnorm* function from The *VGAM* package in R[40]. The number of new data points drawn from each biome was set to the number of observations in the original data set. So, for example, a total of 99 observations were drawn for temperate forests, 41 of which were randomly assigned AM status and 58 EEM status. This data therefore assumes no difference in the biome-specific distributions of AM versus EEM observations, and our analysis should therefore detect no effect of mycorrhizal type on soil C storage. We then modelled soil C content as a function of mycorrhizal type with biome coded as a random effect and counted the number of times the mycorrhizal effect was significant at the 0.05 level (that is, a 1 in 20 random chance). We found that only 6.55% of the 10,000 simulations generated a statistically significant mycorrhizal effect, very slightly more frequently than expected by random chance. Furthermore, they were just as likely to be positive as negative, which means a positive effect of EEM status on soil C was detected less than 5% of the time. In contrast, our analysis finds a positive effect of EEM fungi on soil C storage at the level of $P \leq 0.0001$ (that is, a <1 in 10,000 chance the effect is due to random chance). Therefore, our sampling and data analysis approach did not make us more likely to detect a positive effect of mycorrhizal type on soil C storage.

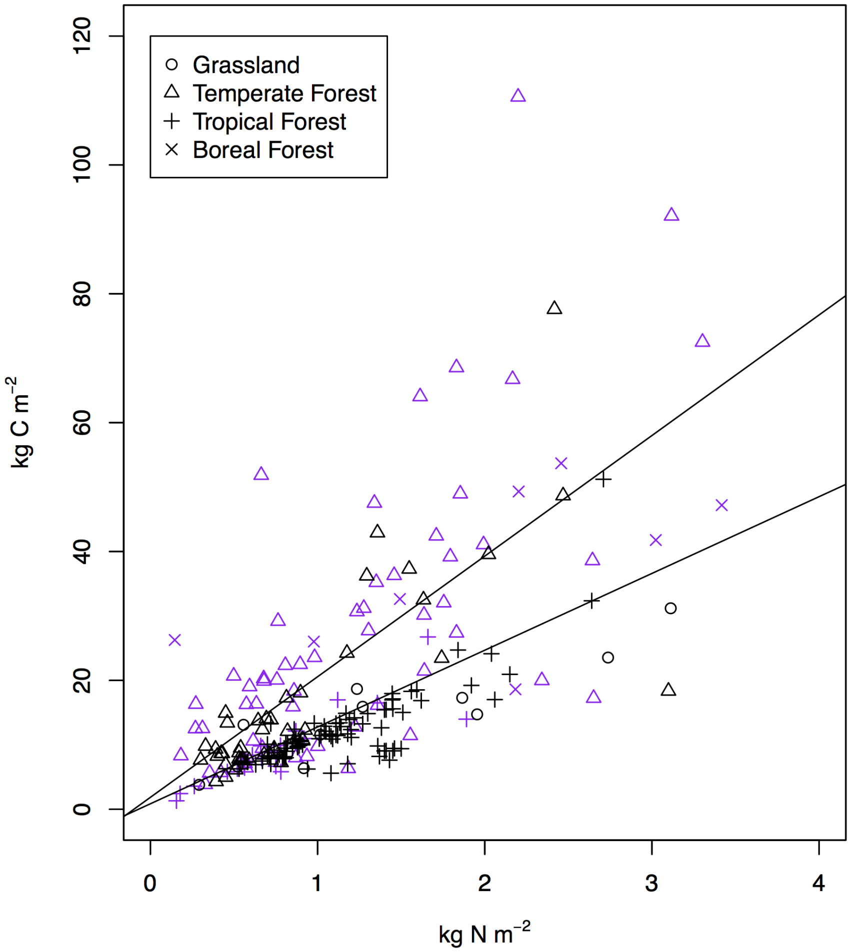
30. Mayes, M. A., Heal, K. R., Brandt, C. C., Phillips, J. R. & Hardine, P. M. Relation between soil order and sorption of dissolved organic carbon in temperate subsoils. *Soil Sci. Soc. Am. J.* **76,** 1027–1037 (2012).
31. Grossman, R. B. & Reinsch, T. G. 2002. in *Methods of Soil Analysis, Part 4: Physical Methods* (eds Dane, J. H. & Topp, C.) 201–203 (Soil Society of America, 2002).
32. Gee, G. W. & Or, D. 2002. in *Methods of Soil Analysis, Part 4: Physical Methods* (eds Dane, J. H. & Topp, C.) 255–293 (Soil Society of America, 2002).
33. Whittaker, R. H. *Communities and Ecosystems* 2nd edn, 111–191 (Macmillan, 1975).
34. Mazerolle, M. J. *AICcmodavg: Model selection and Multimodel Inference based on (Q)AIC(c)* R package version 1.31, http://cran.r-project.org/web/packages/AICcmodavg/index.html (2013).
35. Tofallis, C. Least squares percentage regression. *J. Mod. Appl. Stat. Methods* **7,** 526–534 (2008).
36. Magee, L. R^2 measures based on wald and likelihood ratio joint significance tests. *Am. Stat.* **44,** 250–253 (1990).
37. Aleksandra, M. *lmmfit: Goodness-of-Fit-Measures for Linear Mixed Models with One-Level-Grouping* R package version 1.0, http://cran.r-project.org/web/packages/lmmfit/index.html (2011).
38. Fox, J. & Weisberg, H. S. *An {R} Companion to Applied Regression* 2nd edn (Sage Publications, 2011).
39. Zuur, A. F., Ieno, E. N. & Elphick, C. S. A protocol for data exploration to avoid common statistical problems. *Methods Ecol. Evol.* **1,** 3–14 (2010).
40. Yee, T. W. The VGAM package for categorical data analysis. *J. Stat. Softw.* **32,** 1–34 (2010).



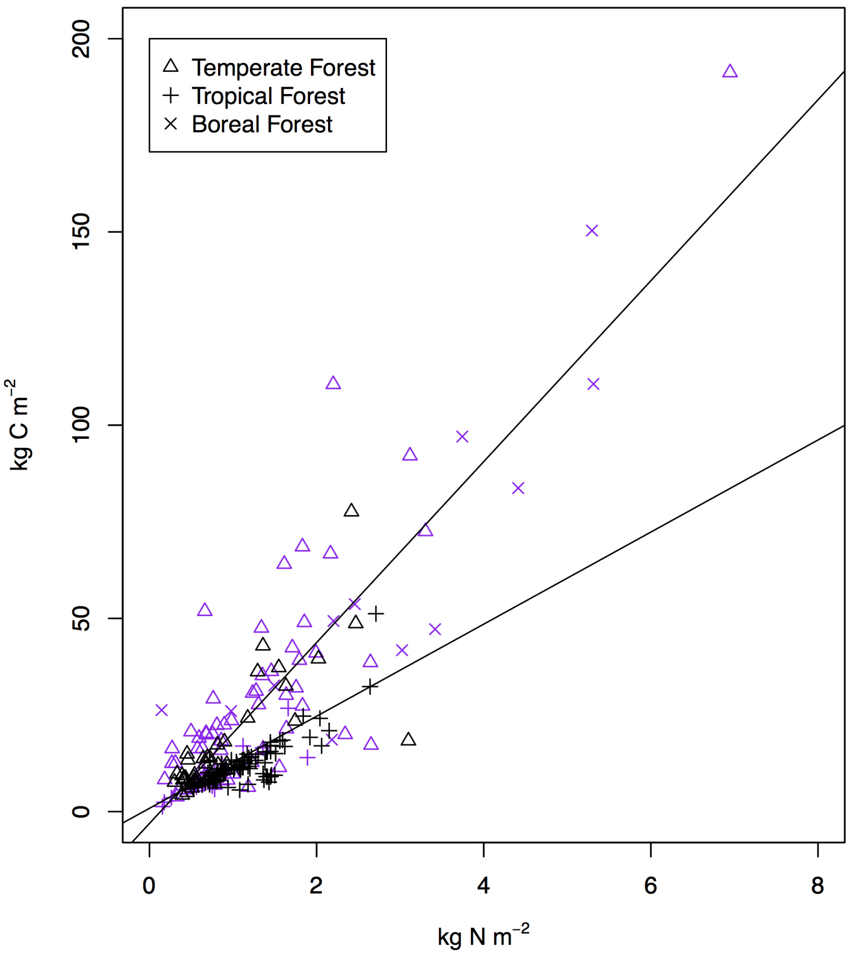
Extended Data Figure 1 | Soil C versus N in the first 50 cm of mineral soil. Purple symbols are EEM observations and black symbols are AM observations. Plotted lines represent univariate regression lines of the respective subsets of the data. We note that plotted lines are univariate regressions of data subsets and are included for visualization purposes only. Removal of the surface organic horizon did not qualitatively change the interpretation of the data. Both the full model and the best AICc-selected model had a significant interactive effect between mycorrhizal type and soil N on soil C storage, with EEM systems storing 1.6 times more C per unit N than AM systems ($P < 0.0001$).



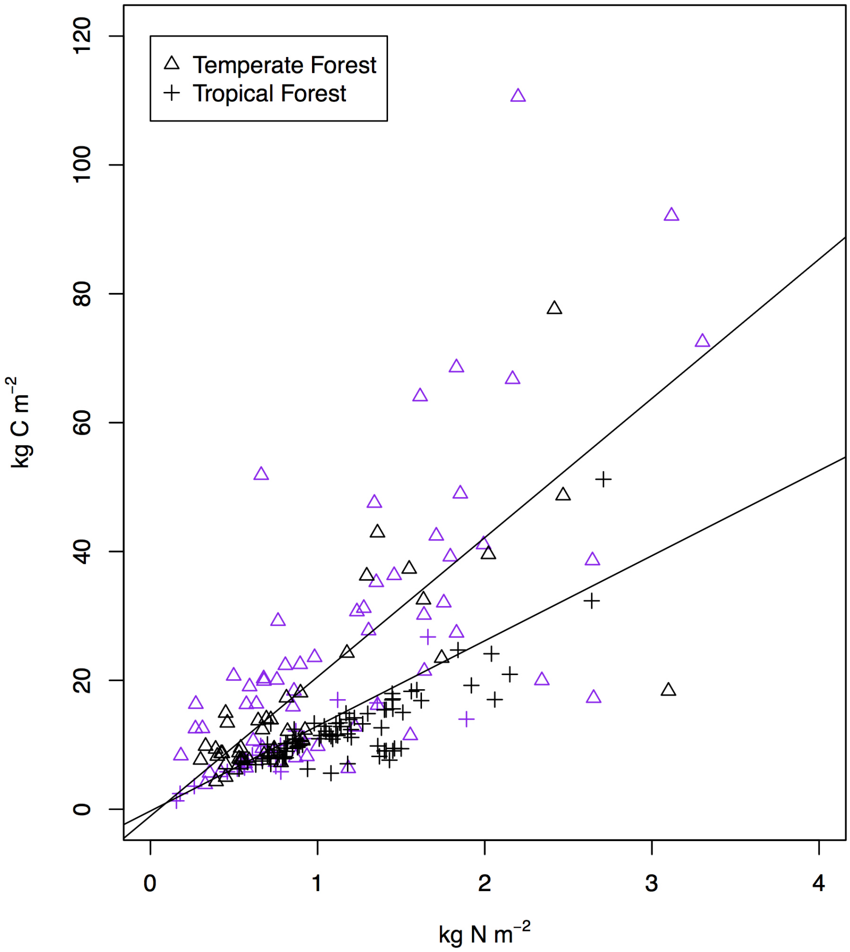
Extended Data Figure 2 | Soil C versus N excluding boreal observations. Purple symbols are EEM observations and black symbols are AM observations. Plotted lines represent univariate regression lines of the respective subsets of the data. We note that plotted lines are univariate regressions of data subsets and are included for visualization purposes only. Both the full model and the best AICc-selected model showed a significant interactive effect of mycorrhizal type and soil N on soil C storage, with EEM systems storing 1.6 times more C per unit N than AM systems ($P = 0.0014$).



Extended Data Figure 3 | Soil C versus N limiting data set to observations with less than 3.5 kg N m^{-2}. Purple symbols are EEM observations and black symbols are AM observations. Plotted lines represent univariate regression lines of the respective subsets of the data. We note that plotted lines are univariate regressions of data subsets and are included for visualization purposes only. Both the full model and the best AICc-selected model found a significant interactive effect of mycorrhizal type and soil N on soil C storage, with EEM systems storing 1.4 times more C per unit N than AM systems ($P = 0.0304$).



Extended Data Figure 4 | Soil C versus N excluding grassland observations. Purple symbols are EEM observations and black symbols are AM observations. Plotted lines represent univariate regression lines of the respective subsets of the data. We note that plotted lines are univariate regressions of data subsets and are included for visualization purposes only. Both the full model and the best AICc-selected model found a significant interactive effect of mycorrhizal type and soil N on soil C storage, with EEM systems storing 1.5 times more C per unit N than AM systems ($P = 0.0023$).



Extended Data Figure 5 | Soil C versus N restricting the analysis to temperate and tropical forest observations only. Purple symbols are EEM and black symbols are AM observations. Plotted lines represent univariate regression lines of the respective subsets of the data. We note that plotted lines are univariate regressions of data subsets and are included for visualization purposes only. Both the full model and the best AICc-selected model incorporated the interactive effect of mycorrhizal type and soil N on soil C storage, with EEM systems storing 1.3 times more C per unit N than AM systems, although the effect was marginally not significant ($P = 0.0690$). We re-emphasize that the full model incorporates biome type, and weights observations by the inverse of their C values, to prevent undue influence of large observations on the estimated effect size.

Extended Data Table 1 | Mineral soil (0–50 cm) analysis regression output from the best AICc-selected model

Parameter	Estimate	Standard error	d.o.f.	*t*-value	*P*-value
(Intercept)	1.239237	0.8238489	218	1.504204	0.1340
EEM	-2.019305	0.8460802	218	-2.386659	0.0179
N	10.362798	0.8404036	218	12.330739	0.0000
EEM:N	5.749893	1.2659686	218	4.541892	0.0000

C ≈ mycorrhizal status × N, random = biome, $R^2_{LR} = 0.091$. C, soil carbon in kg m^{-2}; EEM, the effect of EEM fungi on soil carbon. The *t*-value is from Student's test. d.o.f., degrees of freedom.

Extended Data Table 2 | Removing boreal forests analysis from the best AICc-selected model

Parameter	Estimate	Standard error	d.o.f.	*t*-value	*P*-value
(Intercept)	2.027935	1.71859	209	1.179999	0.2393
EEM	-2.3091	1.594658	209	-1.448022	0.1491
N	9.73505	1.134723	209	8.57923	0.0000
EEM:N	5.580011	1.725335	209	3.23416	0.0014

C ≈ mycorrhizal status × N, random = biome, $R^2_{LR} = 0.89$. C, soil carbon in kg m^{-2}; EEM, the effect of EEM fungi on soil carbon. The *t*-value is from Student's test. d.o.f., degrees of freedom.

Extended Data Table 3 | Restricting range of N content analysis from the best AICc-selected model

Parameter	Estimate	Standard error	d.o.f.	*t*-value	*P*-value
(Intercept)	2.797858	1.915448	215	1.460681	0.1456
EEM	-0.8853	1.560126	215	-0.567454	0.5710
N	9.85441	1.120904	215	8.791487	0.0000
EEM:N	3.615833	1.658742	215	2.179866	0.0304

C ≈ mycorrhizal status × N, random = biome, $R^2_{LR} = 0.83$. C, soil carbon in kg m^{-2}; EEM, the effect of EEM fungi on soil carbon. The *t*-value is from Student's test. d.o.f., degrees of freedom.

Extended Data Table 4 | Removing grasslands analysis from the best AICc-selected model

Parameter	Estimate	Standard error	d.o.f.	*t*-value	*P*-value
(Intercept)	3.745027	2.927873	209	1.279095	0.2023
EEM	-2.215022	1.735368	209	-1.276399	0.2032
N	10.265126	1.371112	209	7.486716	0.0000
EEM:N	5.63454	1.82723	209	3.08365	0.0023

C ≈ mycorrhizal status × N, random = biome, $R^2_{LR} = 0.91$. C, soil carbon in kg m^{-2}; EEM, the effect of EEM fungi on soil carbon. The *t*-value is from Student's test. d.o.f., degrees of freedom.

Extended Data Table 5 | Temperate and tropical biomes only, from the best AICc-selected model

Parameter	Estimate	Standard error	d.o.f.	*t*-value	*P*-value
(Intercept)	1.887213	2.443358	197	0.772385	0.4408
EEM	-0.856856	1.702001	197	-0.50344	0.6152
N	10.357411	1.283155	197	8.071831	0.0000
EEM:N	3.438798	1.881089	197	1.828089	0.0690

C ≈ mycorrhizal status × N, random = biome, $R^2_{LR} = 0.83$. C, soil carbon in kg m^{-2}; EEM, the effect of EEM fungi on soil carbon. The *t*-value is from Student's test. d.o.f., degrees of freedom.

LETTER

OPEN
doi:10.1038/nature12817

The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*)

Juliane C. Dohm[1,2,3]*, André E. Minoche[1,2,3]*, Daniela Holtgräwe[4], Salvador Capella-Gutiérrez[2,3], Falk Zakrzewski[5], Hakim Tafer[6], Oliver Rupp[4], Thomas Rosleff Sörensen[4], Ralf Stracke[4], Richard Reinhardt[7], Alexander Goesmann[4], Thomas Kraft[8], Britta Schulz[9], Peter F. Stadler[6], Thomas Schmidt[5], Toni Gabaldón[2,3,10], Hans Lehrach[1], Bernd Weisshaar[4] & Heinz Himmelbauer[1,2,3]

Sugar beet (*Beta vulgaris* ssp. *vulgaris*) is an important crop of temperate climates which provides nearly 30% of the world's annual sugar production and is a source for bioethanol and animal feed. The species belongs to the order of Caryophylalles, is diploid with $2n = 18$ chromosomes, has an estimated genome size of 714–758 megabases[1] and shares an ancient genome triplication with other eudicot plants[2]. Leafy beets have been cultivated since Roman times, but sugar beet is one of the most recently domesticated crops. It arose in the late eighteenth century when lines accumulating sugar in the storage root were selected from crosses made with chard and fodder beet[3]. Here we present a reference genome sequence for sugar beet as the first non-rosid, non-asterid eudicot genome, advancing comparative genomics and phylogenetic reconstructions. The genome sequence comprises 567 megabases, of which 85% could be assigned to chromosomes. The assembly covers a large proportion of the repetitive sequence content that was estimated[4] to be 63%. We predicted 27,421 protein-coding genes supported by transcript data and annotated them on the basis of sequence homology. Phylogenetic analyses provided evidence for the separation of Caryophyllales before the split of asterids and rosids, and revealed lineage-specific gene family expansions and losses. We sequenced spinach (*Spinacia oleracea*), another Caryophyllales species, and validated features that separate this clade from rosids and asterids. Intraspecific genomic variation was analysed based on the genome sequences of sea beet (*Beta vulgaris* ssp. *maritima*; progenitor of all beet crops) and four additional sugar beet accessions. We identified seven million variant positions in the reference genome, and also large regions of low variability, indicating artificial selection. The sugar beet genome sequence enables the identification of genes affecting agronomically relevant traits, supports molecular breeding and maximizes the plant's potential in energy biotechnology.

During the last 200 years of sugar beet breeding, the sugar content has increased from 8% to 18% in today's cultivars. Breeding has also actively selected for traits like resistance to viral and fungal diseases, improved taproot yield, monogermy of the seed and bolting resistance. After discovering a male sterile cytoplasm, breeders started to develop hybrid varieties and successfully increased yield[5]. Taxonomy assigns *Beta* to the Amaranthaceae family within Caryophylales, an order comprising 11,510 species[6] including cacti, ice plants (Aizoaceae), other drought-tolerant species, and carnivorous plants such as pitcher plants (*Nepenthes*) and sundew (*Drosera*). Until now, no Caryophyllales species have been sequenced.

To provide an extended basis for comparative plant genomics and to support molecular breeding, we sequenced the double haploid sugar beet line KWS2320 as reference genotype, using the Roche/454, Illumina and Sanger sequencing platforms (Extended Data Table 1a, Supplementary Table 1). The initial assembly was integrated with genome-wide genetic and physical map information[2], resulting in 225 genetically anchored scaffolds (394.6 Mb), assigned to nine chromosomes (Table 1, Fig. 1, Extended Data Figs 1 and 2). The chromosomal nomenclature follows a previous study[7] describing a *Beta* karyotype at chromosome arm resolution. The genetically integrated assembly, 'RefBeet', comprised in total 569.0 Mb in 43,721 sequences (2,333 scaffolds and 41,388 unscaffolded contigs) and had an N50 size of 1.7 Mb with 77 scaffolds being of this size or larger. We incorporated Illumina sequencing reads generated from PCR-free libraries and analysed genotyping-by-sequencing data,

Table 1 | Assembly details by chromosome

Chromosome	Total size (Mb)		Number of sequences		N50 size (Mb)	Largest sequence (Mb)
	RefBeet	Incl. GBS data	RefBeet	Incl. GBS data		
1	41.5	51.6	12	28	6.46	8.26
2	39.5	50.4	23	38	2.63	9.20
3	32.3	40.4	17	25	3.06	5.16
4	31.1	53.4	27	56	1.34	4.51
5	56.2	65.4	37	54	2.31	8.59
6	57.8	65.6	31	45	3.38	6.74
7	50.9	54.7	28	42	2.47	10.43
8	40.1	48.0	21	33	2.86	7.39
9	45.2	50.2	29	43	2.29	8.58
	avg. 43.8 sum 394.6	avg. 53.3 sum 479.8	avg. 25 sum 225	avg. 40 sum 364	avg. 2.98	avg. 7.65
un	174.3	86.8	43,496	40,144	0.3	2.02

GBS, genotyping by sequencing.
un, unassigned fraction of the assembly.
avg., average.

[1]Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany. [2]Centre for Genomic Regulation (CRG), C. Dr. Aiguader 88, 08003 Barcelona, Spain. [3]Universitat Pompeu Fabra (UPF), C. Dr. Aiguader 88, 08003 Barcelona, Spain. [4]Bielefeld University, CeBiTec and Department of Biology, Universitätsstraße 25, 33615 Bielefeld, Germany. [5]TU Dresden, Department of Biology, Zellescher Weg 20b, 01217 Dresden, Germany. [6]University of Leipzig, Department of Computer Science, Härtelstraße 16-18, 04107 Leipzig, Germany. [7]Max Planck Genome Centre Cologne, Carl-von-Linné-Weg 10, 50829 Köln, Germany. [8]Syngenta, Box 302, 26123 Landskrona, Sweden. [9]KWS SAAT AG, Grimsehlstraße 31, 37574 Einbeck, Germany. [10]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain.
*These authors contributed equally to this work.

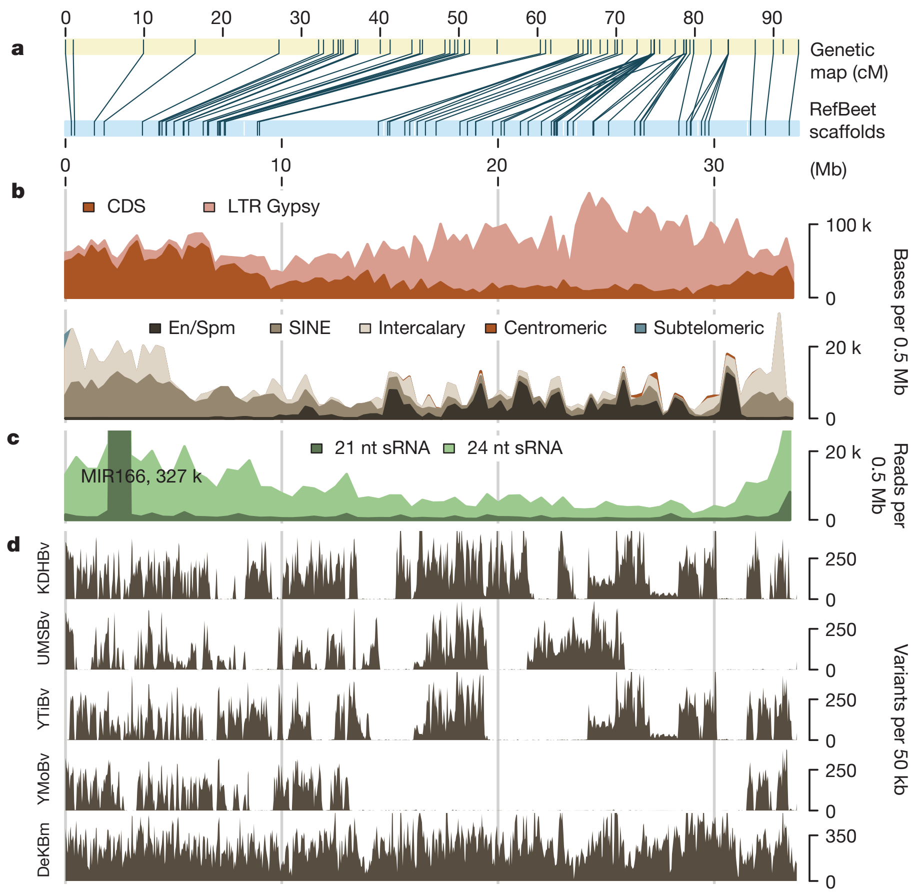


Figure 1 | Genomic features of RefBeet chromosome 1. For chromosomes 2–9 see Extended Data Figs 1 and 2. **a**, Positions of genetic markers in the genetic map[2] and the RefBeet assembly. **b**, Distribution of coding sequence (CDS) and repetitive sequence of the Gypsy type (LTR retrotransposons), the SINE type (non-LTR retrotransposons), the En/Spm type (DNA transposons), and three classes of satellite DNA (intercalary, centromeric, subtelomeric). **c**, Distribution of mapped small RNAs of 21 and 24 nucleotides (nt). The large peak of 21 nt reads (about 327,000 reads mapped) corresponds to the highly expressed microRNA MIR166. **d**, Distribution of genomic variants in four sugar beet accessions and sea beet (DeKBm) compared to RefBeet. Shared and individual low-variation regions per accession are visible (for example, region 30–31 Mb is shared among the sugar beet accessions KDHBv, UMSBv, YTiBv, YMoBv).

leading to an optimized assembly of 566.6 Mb in 2,171 scaffolds and 38,337 unscaffolded contigs. The N50 size was 2.01 Mb (the 72nd scaffold) and the chromosomally assigned fraction 84.7% (Table 1). The assembled part of the genome is assumed to represent the unique regions as well as repetitive regions, which are either short enough to be placed in a unique sequence context or divergent enough to behave as unique entities. A k-mer analysis of Illumina data indicated a genome size of 731 Mb (Extended Data Fig. 3a). We located 94% of publicly available isogenic expressed sequence tags (ESTs) in RefBeet, suggesting that gene-containing regions are comprehensively covered. A sequenced bacterial artificial chromosome (BAC) clone[8] was compared to the corresponding region in RefBeet and found to be correctly assembled within one scaffold. On average, one mismatch and one insertion or deletion (indel) error occurred in 10 kb. RefBeet resolved regions of recombination suppression in centromeric and pericentromeric regions of chromosomes, flanked by regions showing enhanced recombination rates (Extended Data Fig. 4).

We identified 252 Mb (42.3%) of RefBeet as repetitive sequence (Supplementary Data 1). The largest group was long terminal repeat (LTR) retrotransposons (Extended Data Fig. 5a). Gypsy-like elements were enriched in centromeric and pericentromeric regions (Fig. 1, Extended Data Figs 1 and 2). Non-LTR retrotransposons of the long interspersed nuclear element (LINE) type were dispersed, whereas short interspersed nuclear elements (SINEs) were enriched towards chromosome ends. Three major satellite classes were organized in large arrays (Fig. 2). By analysing unassembled genomic data we estimated total amounts of 15.4 Mb centromeric, 6.0 Mb intercalary, and 0.6 Mb subtelomeric satellite DNA, as well as 10.0 Mb of 18S-5.8S-25S and 5S ribosomal genes.

A total of 27,421 protein-coding genes supported by mRNA evidence (Supplementary Table 2) were predicted in RefBeet; 91% included start and stop codons (Supplementary Table 3). The majority of the genes (73.6%) were found within chromosomally assigned scaffolds with on average 5.2 genes per 100 kb, a gene length of 5,252 bp including introns, a coding sequence length of 1,159 bp and 4.9 coding exons per gene. The codon usage was similar to other dicot species (Supplementary Table 4). Homology-based annotation of non-coding RNA genes resulted in 3,005 predictions of tRNAs, microRNAs, small nucleolar RNAs, spliceosomal RNAs and ribosomal RNAs, mainly supported by evidence from isogenic small RNA data (Extended Data Fig. 5b–e, Extended Data Table 1b, Supplementary Table 5).

Based on the translated *Beta vulgaris* gene set and the protein sets of nine other plants (Extended Data Table 1d) we determined 19,747 phylogenetic trees, collectively called 'phylome'[9], and inferred orthologous and paralogous gene relationships (Extended Data Fig. 6a). Previous studies left the phylogeny of rosids, asterids and Caryophyllales unresolved[10] or classified Caryophyllales as a subclade of asterids[11]. A species tree inferred from the collection of gene trees strongly suggested that *Beta vulgaris* branched off before the separation of asterids and rosids (Fig. 3). Thus, according to our data, Caryophyllales represent the most basal eudicot clade among the studied species. The fraction of species-specific genes within eudicots (Fig. 3) was the largest for sugar beet, reflecting its phylogenetic position. The analysis of paralogous genes provided evidence for the absence of a lineage-specific whole genome duplication in *Beta vulgaris* supporting previous studies[2] (Extended Data Fig. 6b–e, Supplementary Table 6).

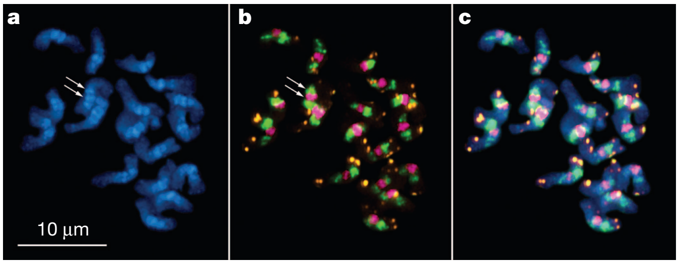


Figure 2 | Fluorescent *in situ* hybridization (FISH) analyses of *Beta vulgaris* chromosomes at early metaphase. **a**, Chromosomes were stained with 4′,6-diamidino-2-phenylindole (DAPI, blue); large blocks of heterochromatin are visible (arrows). **b**, *In situ* hybridization using the major satellites pBV (centromeric, red), pEV (intercalary, green) and pAV (subtelomeric, orange). **c**, Overlayed images of **a** and **b** show the coverage of chromosomes by satellite DNA. Scale bar, 10 μm.

We functionally annotated 17,151 RefBeet genes (63%) based on sequence homology (Supplementary Data 2). The number of disease resistance genes detected was comparatively small, particularly for the STK-domain containing classes (Supplementary Table 7, Supplementary Data 3). In contrast to previous studies[12,13], we found a TNL class resistance gene in the genomes of sugar beet (Bv_22240_ksro) and spinach, both belonging to Amaranthaceae. The phylome tree of Bv_22240_ksro indicated that the presence of a single TNL class gene is a feature of Amaranthaceae, whereas expansion of this gene family is typical for rosids and asterids. The functional categories of expanded and potentially lost gene families (Extended Data Fig. 5f, Supplementary Tables 8, 9) indicate that genes involved in defence and stress compensation represent vital evolutionary targets. The number of transcription factors identified in RefBeet was the lowest of all species studied (Supplementary Table 10). The reference genome sequence enables future experimental approaches to determine if lower gene numbers may alter transcriptional network topologies; Caryophyllales may harbour unknown genes involved in transcriptional control. We identified four sucrose transporter (SUT) orthologues in RefBeet. Phylogenetic analysis including known sucrose transporters suggested a duplication of the *SUT1* gene in Amaranthaceae followed by extensive mutation of one paralogue (Extended Data Fig. 7a). The genome sequences of sugar beet and spinach, both containing the four SUT genes, are an excellent basis for studying the implications of this duplication event.

Previous studies addressing the variation within the genus *Beta* indicated high divergence between genotypes[2,14]. We generated genome sequences of four non-reference sugar beet double haploid accessions (KDHBv, UMSBv, YMoBv, YTiBv) and characterized the genome-wide variation (Extended Data Table 1a, c, Supplementary Tables 3, 11–13). Within RefBeet we identified 7.0 million positions which were variant (77% substituted, 23% deleted) in at least one of the other accessions and 274.9 million positions which were unchanged in all five accessions. We found 2.9 million variants on average per non-reference accession. Coding regions had a prevalence of indels of length three or multiples of three (44%), compared to non-coding regions (16%). The distribution of variants revealed large regions of low variation (Fig. 1, Extended Data Figs 1, 2, 8a–c). Such variation 'deserts' were found in all chromosomes and in all accessions, which might reflect extensive cross-breeding with a limited number of haplotypes in the breeding material, a founder effect, or a bottleneck at the establishment of the crop. However, most of the variation deserts were accession-specific (Extended Data Fig. 8d), probably owing to recombination events that have occurred since the introduction of founder haplotypes into breeding lines. The four accessions shared 50.6 Mb of variation deserts along RefBeet containing 1,824 predicted RefBeet genes (Gene Ontology (GO) term enrichment see Supplementary Table 14). Genes in these regions, analysed in 24 additional sugar beet accessions, showed higher sequence conservation (Extended Data Fig. 8e). These findings suggest that regions of low variation are not maintained by chance, but are rather the result of breeders' selection towards certain genes contained in those regions. The sea beet *Beta maritima* is fully interbreedable with sugar beet and commonly used as a valuable source of resistances against biotic or abiotic stress[15]. We sequenced its genome and identified a total of 75 Mb as variation desert, of which 67 Mb were shared with at least one of the four non-reference *Beta vulgaris* accessions. These regions may represent traces of breeding activities which aimed at introducing sea beet traits into sugar beet. The gene *BvBTC1*, encoded by the *B*-locus[16] and located in a 1.1 Mb RefBeet scaffold on chromosome 2, plays an important role during vernalization. Cultivated lines are homozygotes for the *B* allele resulting in a biennial life cycle. The *B*-locus is located in variation deserts of all five sugar beet lines, whereas the genome of the annual wild form *Beta maritima* shows high variation at this locus, demonstrating that breeding has shaped the genome of sugar beet.

Sugar beet is a hybrid crop based on seed pool lines (male steriles, monogerms) and pollen pool lines (pollinators, multigerms). We identified regions of potentially fixed differences between the two groups: the

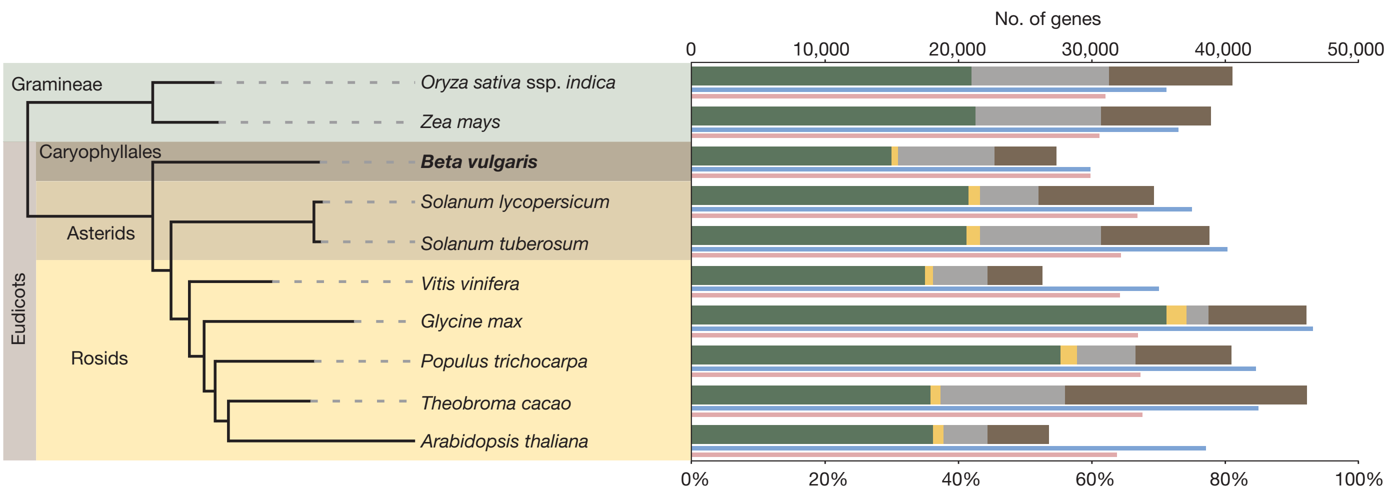


Figure 3 | Phylogenetic relationship of 10 sequenced plant species and comparative gene analysis. Species tree based on maximum-likelihood analysis of a concatenated alignment of 110 widespread single-copy protein sequences (left). The upper bar per species (right with scale on the top) indicates the number of widespread genes that are found in at least 9 of the 10 species (green); eudicot-specific genes that are found in at least 7 of the 8 eudicot species (yellow); species-specific genes with no homologues in other species of this tree (light grey); and remaining genes (brown). The slim bars per species (scale on the bottom) represent the percentage of genes with at least one paralogue (blue) and the percentage of sugar beet genes that have homologues in a given species (pink), respectively.

intersection of shared low-variation regions in seed pool lines and shared high-variation regions of identical variation patterns in pollen pool lines comprised 311 genomic regions (1.6 Mb in total) containing 119 genes.

We performed evidence-based gene predictions in the assemblies of KDHBv, UMSBv, YMoBv and YTiBv. Based on the comparison of 2,112 single copy genes, UMSBv had the largest genetic distance to RefBeet (Extended Data Fig. 7b). The number of accession-specific genes ranged from 79 (RefBeet) to 271 (UMSBv). Genes were analysed for the ratio of non-synonymous to synonymous substitutions, altered start and stop sites, new stop codons, modified splice donor or splice acceptor sites and indels, revealing extensive variation in coding regions (Supplementary Tables 15, 16 and Extended Data Fig. 8f). In addition to allelic variation, the variation in gene content may contribute to heterosis, as has been suggested for maize[17].

The availability of the sugar beet genome sequence very much simplifies fine-mapping of quantitative trait loci and the discovery of causal genes, as single-nucleotide polymorphism (SNP)-based markers can be designed for any region of the genome. Association mapping to identify regions of shared ancestry in sugar beet lines requires at least 100,000 variant positions for genotyping. Such positions can now be selected from a catalogue of seven million variants. The genome sequence facilitates further experimentation to characterize gene functions, which accelerates the identification of rewarding targets for transgenic manipulation, and represents an important foundation for molecular and comparative studies in sugar beet, Caryophyllales and flowering plants. The data presented are key to improvements of the sugar beet crop with respect to yield and quality and towards its application as a sustainable energy crop.

METHODS SUMMARY

Genome sequencing and assembly. Genomic DNA isolated from root and leaf material was sequenced on the Roche/454 FLX, Illumina HiSeq2000 and ABI3730 XL sequencing platforms. The Newbler software was applied on 454, Illumina and Sanger sequencing data to assemble the reference genotype (RefBeet). Contigs of putative bacterial origin and those smaller than 500 bp were removed. Additional lines were sequenced on the HiSeq2000 platform and were assembled using SOAPdenovo. We performed gap-closing and homopolymer error correction using Illumina reads from PCR-based and PCR-free libraries (Extended Data Fig. 3b, c). Chromosome-wise scaffolding using genetic and physical mapping data was assisted by SSPACE (Methods and Supplementary Methods).

Gene annotation. Prediction of protein coding genes was performed using the AUGUSTUS pipeline, with Illumina mRNA-seq reads and other cDNA read data as supporting evidence. Gene models were filtered for transposable element homology. Small and other non-coding RNAs were identified based on homology searches and based on Illumina sequencing data. Repeats were predicted using RepeatModeler, followed by manual curation of the predictions (Methods and Supplementary Methods).

Intraspecific variation. Variant positions (substitutions, indels) were identified by read mapping and scaffold alignment (Methods and Supplementary Methods).

Phylogenetic analysis and species tree reconstruction. The longest protein sequence of each RefBeet gene was used for a Smith–Waterman search against the protein sets of nine other plant species. Alignments were generated and quality-filtered, and phylogenetic trees were calculated for each *Beta vulgaris* sequence. A species tree was generated from a super-tree of all trees and by multi-gene phylogenetic analysis of high-confidence 1:1 orthologues.

Functional annotation. Protein coding gene predictions were functionally annotated based on protein signatures and orthology relationships.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 March; accepted 29 October 2013.

Published online 18 December 2013; corrected online 22 January 2014 (see full-text HTML version for details).

1. Arumuganathan, K. & Earle, E. D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9,** 208–218 (1991).
2. Dohm, J. C. *et al.* Palaeohexaploid ancestry for Caryophyllales inferred from extensive gene-based physical and genetic mapping of the sugar beet genome (*Beta vulgaris*). *Plant J.* **70,** 528–540 (2012).
3. Fischer, H. E. Origin of the 'Weisse Schlesische Rübe' (white Silesian beet) and resynthesis of sugar beet. *Euphytica* **41,** 75–80 (1989).
4. Flavell, R. B., Bennett, M. D., Smith, J. B. & Smith, D. B. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem. Genet.* **12,** 257–269 (1974).
5. Biancardi, E., McGrath, J. M., Panella, L., Lewellen, R. & Stevanato, P. in *Root Tuber Crops* Vol. 7 (ed. Bradshaw, J. E.) 173–219 (Springer, 2010).
6. Stevens, P. *Angiosperm Phylogeny Website* (2012) http://www.mobot.org/MOBOT/research/APweb/.
7. Paesold, S., Borchardt, D., Schmidt, T. & Dechyeva, D. A sugar beet (*Beta vulgaris* L.) reference FISH karyotype for chromosome and chromosome-arm identification, integration of genetic linkage groups and analysis of major repeat family distribution. *Plant J.* **72,** 600–611 (2012).
8. Dohm, J. C., Lange, C., Reinhardt, R. & Himmelbauer, H. Haplotype divergence in *Beta vulgaris* and microsynteny with sequenced plant genomes. *Plant J.* **57,** 14–26 (2009).
9. Huerta-Cepas, J., Dopazo, H., Dopazo, J. & Gabaldón, T. The human phylome. *Genome Biol.* **8,** R109 (2007).
10. The Angiosperm Phylogeny Group. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.* **141,** 399–436 (2003).
11. Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G. & Soltis, D. E. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc. Natl Acad. Sci. USA* (2010).
12. Hunger, S. *et al.* Isolation and linkage analysis of expressed disease-resistance gene analogues of sugar beet (*Beta vulgaris* L.). *Genome* **46,** 70–82 (2003).
13. Tian, Y., Fan, L., Thurau, T., Jung, C. & Cai, D. The absence of TIR-type resistance gene analogues in the sugar beet (*Beta vulgaris* L.) genome. *J. Mol. Evol.* **58,** 40–53 (2004).
14. Schneider, K. *et al.* Analysis of DNA polymorphisms in sugar beet (*Beta vulgaris* L.) and development of an SNP-based map of expressed genes. *Theor. Appl. Genet.* **115,** 601–615 (2007).
15. Biancardi, E., Panella, L. W. & Lewellen, R. T. *Beta maritima: The Origin of Beets* (Springer, 2012).
16. Pin, P. A. *et al.* The role of a pseudo-response regulator gene in life cycle adaptation and domestication of beet. *Curr. Biol.* **22,** 1095–1101 (2012).
17. Schnable, P. S. & Springer, N. M. Progress toward understanding heterosis in crop plants. *Annu. Rev. Plant Biol.* **64,** 71–88 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the BMBF grant "Verbundprojekt GABI BeetSeq: Erstellung einer Referenzsequenz für das Genom der Zuckerrübe (*Beta vulgaris*)", FKZ 0315069A and 0315069B (to H.H. and B.W.) and by the BMBF grant "AnnoBeet: Annotation des Genoms der Zuckerrübe unter Berücksichtigung von Genfunktionen und struktureller Variabilität für Nutzung von Genomdaten in der Pflanzenbiotechnologie.", FKZ 0315962 A, 0315962 B and 0315962 C (to B.W., H.H., and T.S.). We are grateful to M. Zehnsdorf, H. Kang, P. Viehoever, E. Castillo, A. Menoyo and C. Lange for library preparation and sequencing; to D. Datta for sequencing data base calling; to D. Kedra for discussions; and to D. Boyd and M. Isalan for language editing. We thank P. Pin, B. Briggs, and Strube Research for providing plant material and for discussions. We thank Roche for data generation on the 454 sequencing platform (cDNA and genomic 20 kb mate-pairs) and for early access to the Newbler genome assembly software.

Author Contributions H.H., B.W. and J.C.D. conceived the study, H.H., D.H., T.R.S., R.R. and B.W. prepared sequencing data, J.C.D., A.E.M., D.H., S.C.-G., F.Z., H.T., O.R., R.S., A.G., B.S., T.K., P.F.S., T.S. and T.G. designed experiments and analysed the data, H.L. participated in project design, J.C.D., H.H. and A.E.M. wrote the paper with input from all other authors. All authors have read and have approved the manuscript.

Author Information Sequencing raw data (genomic and transcript sequences) have been submitted to the SRA archive with the study accession number SRP023136. The NCBI Bioproject accession is PRJNA41497. The whole-genome shotgun assemblies have been deposited at DDBJ/EMBL/GenBank under the accessions AYZS00000000–AYZY00000000. The GenBank accession numbers KG026656–KG039419 were assigned to BAC end sequences and JY274675 –JY473858 to fosmid end sequences generated in this study. Plant material for *Beta vulgaris* genotype KWS2320 and *Beta maritima* 9W_2101 (DeKBm) are available as seeds by signing a material transfer agreement (MTA). Reprints and permissions information is available at www.nature.com/reprints. A sugar beet website including a genome browser has been set up at http://bvseq.molgen.mpg.de, providing access to assemblies, annotations, gene models and variation data. The sugar beet phylome can be accessed at http://phylomeDB.org. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.H. (Heinz.himmelbauer@crg.es) or B.W. (bernd.weisshaar@uni-bielefeld.de).

doi:10.1038/nature12825

Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease

Carlos Cruchaga[1,2], Celeste M. Karch[1,2]*, Sheng Chih Jin[1]*, Bruno A. Benitez[1], Yefei Cai[1], Rita Guerreiro[3,4], Oscar Harari[1], Joanne Norton[1], John Budde[1], Sarah Bertelsen[1], Amanda T. Jeng[1], Breanna Cooper[1], Tara Skorupa[1], David Carrell[1], Denise Levitch[1], Simon Hsu[1], Jiyoon Choi[1], Mina Ryten[3], UK Brain Expression Consortium (UKBEC)†, Celeste Sassi[3,4], Jose Bras[3], J. Raphael Gibbs[3,4], Dena G. Hernandez[3,4], Michelle K. Lupton[5,6], John Powell[5], Paola Forabosco[7], Perry G. Ridge[8], Christopher D. Corcoran[9,10], JoAnn T. Tschanz[10,11], Maria C. Norton[10,11,12], Ronald G. Munger[12,13], Cameron Schmutz[8], Maegan Leary[8], F. Yesim Demirci[14], Mikhil N. Bamne[14], Xingbin Wang[14], Oscar L. Lopez[15,16], Mary Ganguli[17], Christopher Medway[18], James Turton[18], Jenny Lord[18], Anne Braae[18], Imelda Barber[18], Kristelle Brown[18], The Alzheimer's Research UK (ARUK) Consortium†, Pau Pastor[19,20,21], Oswaldo Lorenzo-Betancor[19], Zoran Brkanac[22], Erick Scott[23], Eric Topol[23], Kevin Morgan[18], Ekaterina Rogaeva[24], Andrew B. Singleton[4], John Hardy[3], M. Ilyas Kamboh[14,15,16], Peter St George-Hyslop[24,25], Nigel Cairns[2,26], John C. Morris[26,27,28], John S. K. Kauwe[8] & Alison M. Goate[1,2,27,28,29]

Genome-wide association studies (GWAS) have identified several risk variants for late-onset Alzheimer's disease (LOAD)[1,2]. These common variants have replicable but small effects on LOAD risk and generally do not have obvious functional effects. Low-frequency coding variants, not detected by GWAS, are predicted to include functional variants with larger effects on risk. To identify low-frequency coding variants with large effects on LOAD risk, we carried out whole-exome sequencing (WES) in 14 large LOAD families and follow-up analyses of the candidate variants in several large LOAD case–control data sets. A rare variant in *PLD3* (phospholipase D3; Val232Met) segregated with disease status in two independent families and doubled risk for Alzheimer's disease in seven independent case–control series with a total of more than 11,000 cases and controls of European descent. Gene-based burden analyses in 4,387 cases and controls of European descent and 302 African American cases and controls, with complete sequence data for *PLD3*, reveal that several variants in this gene increase risk for Alzheimer's disease in both populations. *PLD3* is highly expressed in brain regions that are vulnerable to Alzheimer's disease pathology, including hippocampus and cortex, and is expressed at significantly lower levels in neurons from Alzheimer's disease brains compared to control brains. Overexpression of PLD3 leads to a significant decrease in intracellular amyloid-β precursor protein (APP) and extracellular Aβ42 and Aβ40 (the 42- and 40-residue isoforms of the amyloid-β peptide), and knockdown of PLD3 leads to a significant increase in extracellular Aβ42 and Aβ40. Together, our genetic and functional data indicate that carriers of *PLD3* coding variants have a twofold increased risk for LOAD and that PLD3 influences APP processing. This study provides an example of how densely affected families may help to identify rare variants with large effects on risk for disease or other complex traits.

The identification of pathogenic mutations in *APP*, presenilin 1 (*PSEN1*) and *PSEN2*, and the association of apolipoprotein E (*APOE*) genotype with disease risk led to a better understanding of the pathobiology of Alzheimer's disease, and the development of novel animal models and therapies for this disease[3]. Recent studies using next-generation sequencing have also identified a protective variant in *APP*[4], and a low-frequency variant in *TREM2* associated with Alzheimer's disease risk[5–8] with odds ratio close to that of one *APOE4* allele. These studies have led to the identification of functional variants with large effects on Alzheimer's disease pathogenesis, in contrast to the loci identified through GWAS[1,2]. Low-frequency coding variants not detected by GWAS may be a source of functional variants with a large effect on LOAD risk[5–8]; however, the identification of such variants remains challenging because most study designs require WES in very large data sets. One potential solution is to perform WES or whole-genome-sequencing in a highly selected population at increased risk for disease followed by a combination of genotyping and deep re-sequencing of the variant or gene of interest in large numbers of cases and controls.

We reported previously that families with a clinical history of LOAD in four or more individuals are enriched for genetic risk variants in known Alzheimer's disease and frontotemporal dementia (FTD) genes, but some of these families do not carry pathogenic mutations in the known Alzheimer's disease or FTD genes[9,10], suggesting that additional genes may contribute to LOAD risk. We ranked 868 LOAD families from the National Institute on Aging (NIA)-LOAD study based on number of affected individuals, number of generations affected, the number of

[1]Department of Psychiatry, Washington University, 425 South Euclid Avenue, St. Louis, Missouri 63110, USA. [2]Hope Center Program on Protein Aggregation and Neurodegeneration, Washington University 425 South Euclid Avenue, St. Louis, Missouri 63110, USA. [3]Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK. [4]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Building 35 Room 1A1014, 35 Lincoln Drive, Bethesda, Maryland 20892, USA. [5]Institute of Psychiatry, King's College London, 16 De Crespigny Park, London SE5 8AF, UK. [6]Neuroimaging Genetics, QIMR Berghofer Medical Research Institute, 300 Herston Road, Herston, Queensland 4006, Australia. [7]Istituto di Genetica delle Popolazioni – CNR, Trav. La Crucca, 3 - Reg. Baldinca - 07100 Li Punti, Sassari, Italy. [8]Department of Biology, Brigham Young University, Provo, Utah 84602, USA. [9]Department of Mathematics and Statistics, Utah State University, Logan, Utah 84322, USA. [10]Center for Epidemiologic Studies, Utah State University, Logan, Utah 84322, USA. [11]Department of Psychology, Utah State University, Logan, Utah 84322, USA. [12]Department of Family Consumer and Human Development, Utah State University, Logan, Utah 84322, USA. [13]Department of Nutrition, Dietetics, and Food Sciences, Utah State University, Logan, Utah 84322, USA. [14]Department of Human Genetics, University of Pittsburgh, 130 Desoto Street, Pittsburgh, Pennsylvania 15261, USA. [15]Alzheimer's Disease Research Center, University of Pittsburgh, 130 Desoto Street, Pittsburgh, Pennsylvania 15261, USA. [16]Department of Neurology, University of Pittsburgh, 130 Desoto Street, Pittsburgh, Pennsylvania 15261, USA. [17]Department of Psychiatry, University of Pittsburgh, 130 Desoto Street, Pittsburgh, Pennsylvania 15261, USA. [18]Human Genetics, School of Molecular Medical Sciences, University of Nottingham, Queen's Medical Centre, Nottingham NG7 2UH, UK. [19]Neurogenetics Laboratory, Division of Neurosciences, Center for Applied Medical Research, University of Navarra, Avenida Pío XII, 55. 31008 Pamplona, Navarra, Spain. [20]Department of Neurology, Clínica Universidad de Navarra, School of Medicine, University of Navarra Avenida Pío XII, 36. 31008 Pamplona, Spain. [21]CIBERNED, Centro de Investigación Biomédica en Red de Enfermedades Neurodegenerativas, Instituto de Salud Carlos III, Spain. [22]University of Washington, 325 Ninth Avenue, Seattle, Washington 98104-2499, USA. [23]The Scripps Research Institute, La Jolla, California 3344 North Torrey Pines Court, La Jolla, California 92037, USA. [24]Tanz Centre for Research in Neurodegenerative Diseases, University of Toronto, 60 Leonard Avenue, Toronto, Ontario M5T 2S8, Canada. [25]Cambridge Institute for Medical Research, and the Department of Clinical Neurosciences, University of Cambridge, Hills Road, Cambridge CB2 0XY, UK. [26]Pathology and Immunology, Washington University, 425 South Euclid Avenue, St. Louis, Missouri 63110, USA. [27]Department of Neurology, Washington University, 425 South Euclid Avenue, St. Louis, Missouri 63110, USA. [28]Knight ADRC, Washington University, 425 South Euclid Avenue, St. Louis, Missouri 63110, USA. [29]Department of Genetics, Washington University, 425 South Euclid Avenue, St. Louis, Missouri 63110, USA.
*These authors contributed equally to this work.
†A list of authors and affiliations appears at the end of the paper.

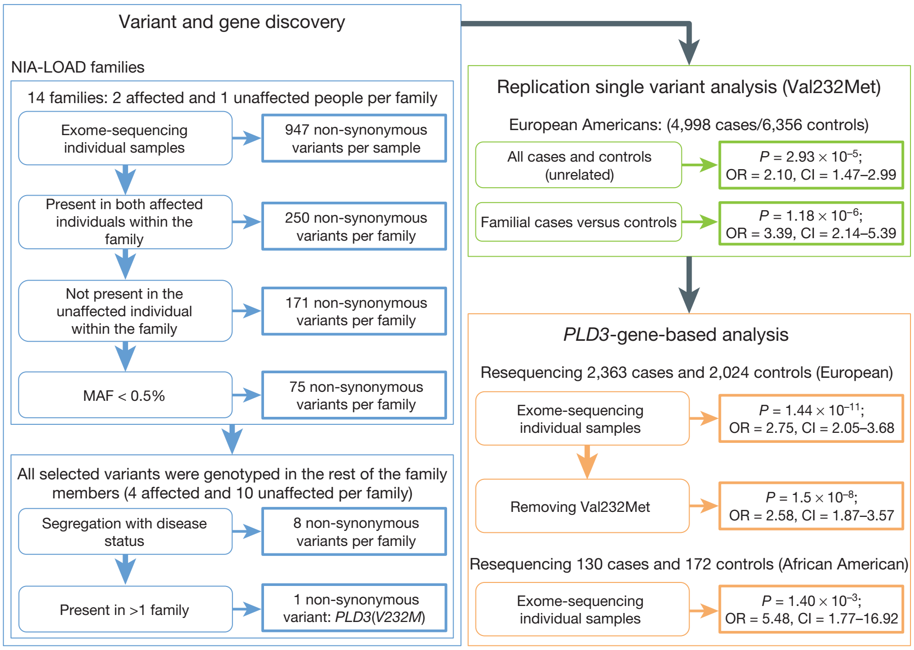


Figure 1 | Summary of the main genetic findings. The diagram shows the steps used to filter the variants identified by exome-sequencing, which led to the identification of the *PLD3*(*V232M*) variant. The diagram also shows the subsequent genetic analyses in large case–control data sets that validated the association of the Val232Met variant and *PLD3* with risk for Alzheimer's disease. CI, confidence interval; OR, odds ratio.

affected and unaffected individuals with DNA available, the number of individuals with a definite or probable diagnosis of Alzheimer's disease, early age at onset (AAO) and *APOE* genotype (discarding families in which *APOE4* segregates with disease status), and 14 were selected to perform WES. In the 14 selected families, there were at least four affected individuals per family, with DNA available for at least three of these individuals. We sequenced at least two affected individuals per family, prioritizing distantly related affected individuals with the earliest AAO. We also sequenced one unaffected individual in nine families and two unaffected individuals in one family. In total, we performed WES on 29 affected individuals and 11 unaffected individuals from 14 families of European American ancestry (Supplementary Table 1 and Supplementary Fig. 2).

All variants shared by affected individuals but absent in unaffected individuals within a family, with a minor allele frequency (MAF) lower than 0.5% in the Exome Variant Server (EVS; http://evs.gs.washington.edu/EVS/) were selected and genotyped in the remaining family members to determine segregation with disease (Supplementary Information). We next examined whether individual variants or variants in the same gene segregated with disease in more than one family. A single variant, rs145999145 (Val232Met, *PLD3*, chromosome 19q13.2), segregated with disease in two independent families (Fig. 1 and Supplementary Fig. 1). We then sought to determine whether this variant was associated with increased risk for sporadic Alzheimer's disease in seven independent data sets (4,998 Alzheimer's disease cases and 6,356 controls of European descent from the Knight Alzheimer's Disease Research Centre (ADRC), NIA-LOAD, NIA-UK data set, Cache-County study, the Universities of Toronto, Nottingham and Pittsburgh, the National Institutes of Mental Health (NIMH) Alzheimer's disease series, and the Wellderly study[7,11–14]; Extended Data Table 1). *PLD3*(*V232M*) was associated with both Alzheimer's disease risk ($P = 2.93 \times 10^{-05}$, odds ratio = 2.10, 95% CI = 1.47–2.99; Table 1) and AAO ($P = 3 \times 10^{-3}$; Extended Data Fig. 1). The frequency of *PLD3*(*V232M*) was higher in Alzheimer's disease cases compared to controls in each age–gender–ethnicity matched data set, with a similar estimated odds ratio for each data set (Extended Data Table 1 and Extended Data Fig. 2), suggesting that the association is unlikely to be a false positive due to population stratification. This was confirmed when population principal components derived from GWAS data were included (Supplementary Information, and Supplementary Figs 2 and 3). The association of the Val232Met variant with Alzheimer's disease risk was also independent of *APOE* genotype (Supplementary Information, Supplementary Table 3 and Supplementary Fig. 4).

LOAD risk variants, such as *APOE4*, are most common in Alzheimer's disease cases with a family history of disease and least common in elderly controls without disease[8,9]. We examined the frequency of Val232Met in three groups of elderly individuals without dementia stratified by age (>65 years, >70 years and >80 years; Table 1) and compared them with

Table 1 | Association between *PLD3*(*V232M*) and Alzheimer's disease risk in individuals of European descent.

Group		Count (carriers/non-carriers)	Frequency (%)	Odds ratio (95% CI)	***P* value**
Control group	All controls	50/6,306	0.79	NA	NA
	>65 years, no dementia	9/1,690	0.52	NA	NA
	>70 years, no dementia	5/1,248	0.39	NA	NA
	>80 years, no dementia	1/375	0.26	NA	NA
Cases group	All Alzheimer's disease cases	82/4,916	1.64	*2.10 (1.47–2.99)	2.93×10^{-5}
				†3.13 (1.57–6.24)	3.54×10^{-4}
				‡4.16 (1.68–10.29)	2.34×10^{-4}
	Index cases (families)	29/1,077	2.62	*3.39 (2.14–5.39)	1.18×10^{-6}
				†5.05 (2.38–10.41)	5.14×10^{-6}
				‡6.72 (2.59–17.52)	5.23×10^{-6}
	Sporadic Alzheimer's disease cases	53/3,839	1.36	*1.74 (1.18–2.57)	5.70×10^{-3}
				†2.59 (1.27–5.26)	5.20×10^{-3}
				‡3.44 (1.37–8.63)	3.20×10^{-3}

The table shows the counts for minor allele carriers and non-carriers. *P* values were calculated using Fisher's exact test. Only individuals of European descent were included in this analysis. The carrier frequency for the Val232Met variant in the Exome Variant Server (EVS) is 0.99%. *Odds ratio and *P* value in comparison with all controls. †Odds ratio and *P* value in comparison with individuals aged over 65 years who do not have dementia. ‡Odds ratio and *P* value in comparison with individuals aged over 70 years who do not have dementia. NA, not applicable.

sporadic versus familial Alzheimer's disease cases. As predicted for an Alzheimer's disease risk allele, Val232Met showed age-dependent differences in frequency among controls with the lowest frequency in the Wellderly data set, a series composed of healthy individuals without dementia, who were older than 80 years (carrier frequency 0.27%). Similarly, no Val232Met carriers were found among the 303 individuals without dementia who had normal cerebrospinal fluid Aβ42 and tau profiles, suggesting that the calculated odds ratio for the Val232Met variant when compared to all controls may be an underestimation (Supplementary Information and Supplementary Table 4). As proposed, the frequency of Val232Met was higher in familial cases than in sporadic cases (2.62% in familial versus 1.36% in sporadic cases).

Several risk variants have been observed in *APP*, *PSEN1* and *PSEN2* and *APOE*, supporting the role of these genes in Alzheimer's disease risk[3,4]. To identify additional risk variants in *PLD3*, we sequenced the *PLD3* coding region in 2,363 cases and 2,024 controls of European descent (Extended Data Tables 2 and 3). Fourteen variants were observed more frequently in cases than in controls, including nine variants that were unique to cases (Fig. 2a and Supplementary Information). The gene-based burden analysis resulted in a genome-wide significant association of carriers of *PLD3* coding variants among Alzheimer's disease cases (7.99%) compared to controls (3.06%; $P = 1.44 \times 10^{-11}$; odds ratio = 2.75, 95% CI = 2.05–3.68). When the Val232Met variant was excluded, the association remained highly significant, still passing genome-wide multiple-test correction ($P = 1.58 \times 10^{-8}$; odds ratio = 2.58, 95% CI = 1.87–3.57; Extended Data Table 3), indicating that there are additional variants in *PLD3* that increase risk for Alzheimer's disease independent of Val232Met. There were two additional highly conserved variants (Supplementary Fig. 5), that were nominally associated with LOAD risk: Met6Arg ($P = 0.02$; odds ratio = 7.73, 95% CI = 1.09–61), and Ala442Ala ($P = 3.78 \times 10^{-7}$; odds ratio = 2.12, 95% CI = 1.58–2.83). The Ala442Ala variant showed an association with LOAD risk in four independent series (Extended Data Table 4). This variant was included in the gene-based analysis because our bioinformatic and functional analyses indicate that this variant affects splicing and gene expression (see below).

If the association of *PLD3* with Alzheimer's disease risk is real, it is possible that rare coding variants in *PLD3* in other populations will also increase risk for Alzheimer's disease. We therefore sequenced *PLD3* in 302 African American Alzheimer's disease cases and controls. Both the Val232Met and the Ala442Ala variants were found in Alzheimer's disease cases but not controls, and the Ala442Ala variant showed a significant association with Alzheimer's disease risk ($P = 0.03$). There was also a significant association with LOAD risk at the gene level ($P = 1.4 \times 10^{-3}$; odds ratio = 5.48, 95% CI = 1.77–16.92; Fig. 1, Extended Data Table 5 and Supplementary Information). This consistent evidence of association with Alzheimer's disease risk, at the single-nucleotide polymorphism (SNP) and gene level in two different populations strongly supports *PLD3* as an Alzheimer's disease risk gene.

To begin to understand the link between *PLD3* and Alzheimer's disease, we analysed *PLD3* expression in Alzheimer's disease case and control brains. In human brain tissue from cognitively normal individuals, *PLD3* showed high levels of expression in the frontal, temporal and occipital cortices and hippocampus (Supplementary Fig. 6). Using data from gene expression in laser-captured neurons from Alzheimer's disease cases and controls, *PLD3* gene expression was significantly lower in Alzheimer's disease cases compared to controls ($P = 8.10 \times 10^{-10}$; Fig. 2b). This result was replicated in three additional independent data sets (Supplementary Information and Extended Data Fig. 3). Bioinformatic analyses predicted that the Ala442Ala variant affects alternative splicing (Supplementary Fig. 7 and Supplementary Information). We found that Ala442Ala is associated with lower levels of total *PLD3* messenger RNA (Fig. 2D) and lower levels of transcripts containing exon 11 (Fig. 2c and Supplementary Fig. 8), supporting the functional effect of this variant.

PLD3 is a non-classical, poorly characterized member of the PLD superfamily of phospholipases. PLD1 and PLD2 have been previously implicated in APP trafficking and Alzheimer's disease[15–17]. To determine

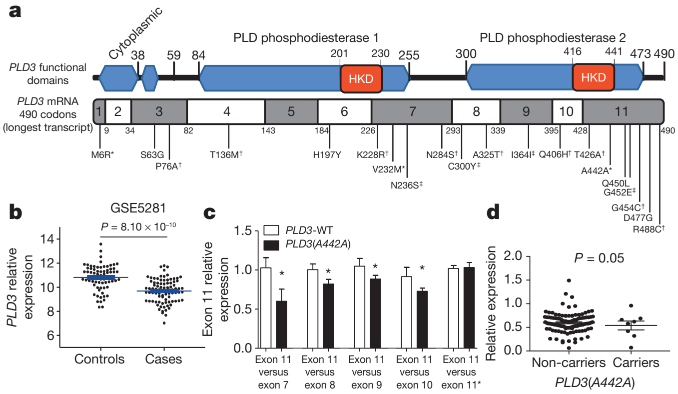


Figure 2 | Most of the *PLD3* coding variants are located in exon 11, and the Ala442Ala variant affects splicing. **a**, Schematic representation of *PLD3* and the relative position of the *PLD3* variants. PLD3 has two PLD phosphodiesterase domains, which contain an HKD signature motif (H-X-K-X(4)-D-X(6)-G-T-X-N, where X represents any amino acid residue). The scheme also shows the exon composition of the longest *PLD3* mRNA and the position of the variants found in this study. *Variants significantly associated with Alzheimer's disease risk. †Variants found only in Alzheimer's disease cases. ‡Variants that are more frequent in Alzheimer's disease cases than in controls. **b**, *PLD3* neuronal gene expression is significantly lower in Alzheimer's disease cases compared to controls. We used the Gene Expression Omnibus data set GSE5281 (ref. 26), in which neurons were laser-captured to analyse whether *PLD3* mRNA expression levels are different between Alzheimer's disease cases and cognitively normal elderly individuals. **c**, **d**, The *PLD3*(*A442A*) variant is associated with lower total *PLD3* mRNA expression and lower levels of exon11 containing transcripts. Primers specific to exons 7, to 11 (two pairs of primers) were designed with PrimerExpress (**c**). cDNA from 8 *PLD3*(*A442A*) carriers and 10 age-, gender-, *APOE*-, clinical dementia rating (CDR)- and post-mortem interval (PMI)-matched individuals were extracted from parietal lobe. Relative expression of exon 11 compared to the other exons was calculated by the ΔCt (changes in cycle threshold) method. Exon-11-containing transcripts were 20% lower in Ala442Ala carriers ($P < 0.05$) in comparison to exon-7–10-containing transcripts. Graphs represent the mean ± s.e.m. Real-time PCR was used to quantify total *PLD3* mRNA and standardized using *GADPH* mRNA as a reference (**d**). *P* value in **d** is for the gene-expression levels of major allele carriers versus minor allele carriers after correcting for dementia severity.

whether PLD3 also affects APP processing, wild-type human PLD3 was overexpressed in mouse neuroblastoma (N2A) cells that stably express wild-type human APP695 (*APP695-WT*; cells termed N2A-695). In this system extracellular Aβ42 and Aβ40 were decreased by 48% and 58%, respectively, compared to the empty vector ($P < 0.0001$; Fig. 3a). Conversely, knockdown of endogenous *PLD3* expression by short hairpin RNA (shRNA) in N2A-695 cells resulted in higher levels of extracellular Aβ42 and Aβ40 than in cells transfected with scrambled shRNA (Fig. 3b). To determine whether the observed effects on APP processing were unique to PLD3 or common among the phospholipase D protein family, we co-expressed APP695-WT with PLD1, PLD2 and PLD3 in human embryonic kidney (HEK293T) cells. Overexpression of PLD3, but not empty vector, PLD1 or PLD2, resulted in a substantial decrease in full-length APP levels (Fig. 3c). Extracellular Aβ42 and Aβ40 levels were significantly reduced in cells overexpressing PLD1, PLD2 and PLD3 compared to control (Fig. 3c). Interestingly, overexpression of catalytically inactive PLD1 and PLD2 variants (*PLD1*(*K898R*) and *PLD2*(*K758R*)) restored extracellular Aβ42 and Aβ40 levels to control values, demonstrating that this is in part a phospholipase-activity-dependent effect (Fig. 3c). Overexpression of a PLD3 dominant-negative variant (*PLD3*(*K418R*)) that inhibits myotube formation[18] failed to restore full-length APP and Aβ42 and Aβ40 to normal levels (Fig. 3c). Furthermore, PLD3 can be co-immunoprecipitated with APP in cultured cells (Extended Data Fig. 4). Together, these studies demonstrate that PLD3 has a role in APP processing that is functionally distinct from PLD1 and PLD2. These findings are consistent with the human

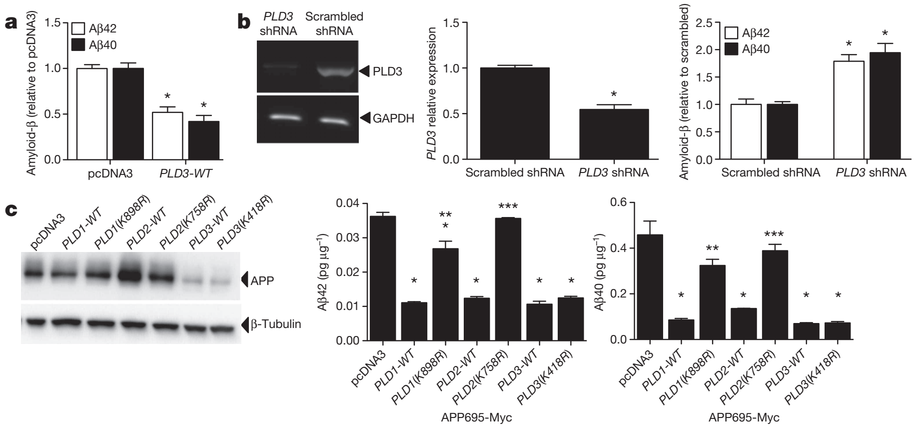


Figure 3 | PLD3 affects APP processing. **a**, **b**, Overexpression and knockdown of PLD3 produce opposing effects on extracellular amyloid-β levels. N2A cells stably expressing human APP695-WT were transiently transfected with vectors containing no insert (*pcDNA3*), human *PLD3-WT*, scrambled shRNA (Origene), or mouse *PLD3* shRNA (Origene) for 48 h. Cell media were analysed with Aβ40 and Aβ42 ELISAs and corrected for total intracellular protein. Amyloid-β levels were then expressed relative to pcDNA3. Graphs represent the mean ± s.e.m. Overexpression of human PLD3 produces significantly less extracellular Aβ42 and Aβ40 (**a**). $*P < 0.0001$. Knockdown of endogenous PLD3 cells produces significantly more extracellular Aβ42 and Aβ40 (**b**). $*P < 0.002$. **c**, Members of the PLD protein family have different effects on APP processing. HEK293T cells were transiently transfected with vectors containing human *APP-WT* and an empty vector (*pcDNA3*), *PLD1, PLD2* or *PLD3-WT*, or *PLD1, PLD2, PLD3* carrying a dominant-negative mutation. Left panel, PLD3 affects full-length APP levels. Cell lysates were extracted in non-ionic detergent, analysed by SDS–PAGE and immunoblot with antibodies to the Myc-tag on APP (9E10) or β-tubulin. Middle (Aβ42) and right (Aβ40) panels, cell media were analysed with Aβ40 and Aβ42 ELISAs and corrected for total intracellular protein. Graphs represent the mean ± s.e.m. $*P < 0.01$, different from pcDNA3; $**P = 0.002$, different from PLD1-WT; $***P < 0.0001$, different from PLD2-WT. Images are representative of at least three replicate experiments.

genetic and brain expression data presented above; lower PLD3 expression and function is correlated with higher APP and amyloid-β levels and with more extensive Alzheimer's-disease-specific pathology (Supplementary Table 4).

Here we provide extensive genetic evidence that *PLD3* is an Alzheimer's disease risk gene: genome-wide significant evidence that rare variants in *PLD3* increase risk for Alzheimer's disease in multiple data sets and two populations. In addition, our functional studies confirm that PLD3 affects APP processing, in a manner that is consistent with increased risk for Alzheimer's disease[3,19]. This work also provides a second example of a novel gene containing rare variants that influence risk for Alzheimer's disease[5,7,8]. Although these variants have low population attributable fraction (proportion of cases in the population attributable to *PLD3* variants) and diagnostic utility owing to their rarity, they provide important and novel insights into Alzheimer's disease pathogenesis. Our success in identifying multiple families carrying the Val232Met variant and the enrichment of this variant in LOAD families compared to sporadic Alzheimer's disease cases demonstrates the power of using a highly selected sample of multiplex LOAD families for variant discovery. The studies on *TREM2* (refs 5–8), and this report, suggest that next-generation sequencing projects will identify additional low-frequency and rare variants associated with Alzheimer's disease.

METHODS SUMMARY

Participants. Samples were obtained from seven independent data sets totalling 4,998 Alzheimer's disease cases and 6,356 controls of European descent from the Knight ADRC, NIA-LOAD, NIA-UK data set, Cache-County study, the Universities of Toronto, Nottingham and Pittsburgh, the NIMH-AD series, and the Wellderly study[7,11–14].

Exome sequencing. Enrichment of coding exons and flanking intronic regions was carried out using a solution hybrid selection method with the SureSelect human all exon 50-Mb kit (Agilent Technologies) as previously described[20].

SNP genotyping. SNPs were genotyped using the Illumina Golden Gate, Sequenom, KASPar[21,22] and/or Taqman.

***PLD3* sequencing.** *PLD3* was sequenced using a pooled-DNA sequencing design as described previously[9,23,24]. All rare missense or splice site variants were then validated by Sequenom and KASPar genotyping.

Gene-expression and alternative splicing analyses. Total RNA was extracted using the RNeasy mini kit (Qiagen). Complementary DNA was prepared from the total RNA, using the High-Capacity cDNA Archive kit (ABI). Gene-expression levels were analysed by real-time polymerase chain reaction (PCR), using an ABI-7900 real-time PCR system.

Statistical analyses. All of the single SNP analyses were performed using a Fisher's exact test. Allelic association with risk for Alzheimer's disease was tested using 'proc logistic' in SAS, including *APOE* genotype, age, principal component (PC) factors, from population stratification analyses and study as covariates when available. Gene-based analyses were performed using the optimal SNP-set Kernel Association Test (SKAT-O)[25].

Cell-based studies. To assess the effects of PLD3 expression on APP cleavage, vectors containing *PLD3-WT* or *PLD3* shRNA were transiently transfected in mouse N2A cells stably expressing human *APP695-WT*. Aβ40 and Aβ42 were measured in conditioned media by enzyme-linked immunosorbant assay (ELISA) (Invitrogen). *PLD3* silencing was confirmed by quantitative PCR (qPCR). To assess the effects of PLD proteins on APP cleavage, HEK293T cells were transiently transfected with vectors containing *PLD1*, *PLD2* and *PLD3-WT* or dominant-negative mutations. Aβ40 and Aβ42 were measured in conditioned media by ELISA. Full-length APP levels were measured by immunoblot analysis of cell lysates.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 August; accepted 31 October 2013.
Published online 11 December 2013.

1. Bertram, L., McQueen, M., Mullin, K., Blacker, D. & Tanzi, R. The AlzGene Database. Alzheimer Research Forum. http://www.alzgene.org (26 January 2013).
2. Lambert, J. C. *et al.* Extended meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genet.* http://dx.doi.org/10.1038/ng.2802 (27 October 2013).
3. Goate, A. & Hardy, J. Twenty years of Alzheimer's disease-causing mutations. *J. Neurochem.* **120** (Suppl. 1), 3–8 (2012).
4. Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488,** 96–99 (2012).
5. Benitez, B. A. *et al. TREM2* is associated with the risk of Alzheimer's disease in Spanish population. *Neurobiol. Aging* **34,** e1715–e31717 (2013).
6. Benitez, B. A. & Cruchaga, C. *TREM2* and neurodegenerative disease. *N. Engl. J. Med.* **369,** 1567–1568 (2013).
7. Guerreiro, R. *et al. TREM2* variants in Alzheimer's disease. *N. Engl. J. Med.* **368,** 117–127 (2013).
8. Jonsson, T. *et al.* Variant of *TREM2* associated with the risk of Alzheimer's disease. *N. Engl. J. Med.* (2013).
9. Cruchaga, C. *et al.* Rare variants in *APP*, *PSEN1* and *PSEN2* increase risk for AD in late-onset Alzheimer's disease families. *PLoS ONE* **7,** e31039 (2012); correction http://dx.doi.org/10.1371/annotation/c92e16da-7733-421d-b063-1db19488daa6 (2012).

10. Harms, M. *et al.* *C9orf72* hexanucleotide repeat expansions in clinical alzheimer disease. *JAMA Neurol.* **70,** 736–741 (2013).
11. Cruchaga, C. *et al.* GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's Disease. *Neuron* (2013).
12. Wijsman, E. M. *et al.* Genome-wide association of familial late-onset Alzheimer's disease replicates *BIN1* and *CLU* and nominates *CUGBP2* in interaction with *APOE*. *PLoS Genet.* **7,** e1001308 (2011).
13. Breitner, J. C. *et al.* APOE-ε4 count predicts age when prevalence of AD increases, then declines: the Cache County Study. *Neurology* **53,** 321–331 (1999).
14. Kamboh, M. I. *et al.* Genome-wide association study of Alzheimer's disease. *Transl. Psychiatry* **2,** e117 (2012).
15. Cai, D. *et al.* Phospholipase D1 corrects impaired βAPP trafficking and neurite outgrowth in familial Alzheimer's disease-linked presenilin-1 mutant neurons. *Proc. Natl Acad. Sci. USA* **103,** 1936–1940 (2006).
16. Cai, D. *et al.* Presenilin-1 uses phospholipase D1 as a negative regulator of β-amyloid formation. *Proc. Natl Acad. Sci. USA* **103,** 1941–1946 (2006).
17. Oliveira, T. G. *et al.* Phospholipase d2 ablation ameliorates Alzheimer's disease-linked synaptic dysfunction and cognitive deficits. *J. Neurosci.* **30,** 16419–16428 (2010).
18. Osisami, M., Ali, W. & Frohman, M. A. A role for phospholipase D3 in myotube formation. *PLoS ONE* **7,** e33341 (2012).
19. Hardy, J. & Allsop, D. Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends Pharmacol. Sci.* **12,** 383–388 (1991).
20. Benitez, B. A. *et al.* Exome-sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis. *PLoS ONE* **6,** e26741 (2011).
21. Cruchaga, C. *et al.* Association of TMEM106B gene polymorphism with age at onset in granulin mutation carriers and plasma granulin protein levels. *Arch. Neurol.* **68,** 581–586 (2011).
22. Cruchaga, C. *et al.* Association and expression analyses with single-nucleotide polymorphisms in *TOMM40* in Alzheimer disease. *Arch. Neurol.* **68,** 1013–1019 (2011).
23. Jin, S. C. *et al.* Pooled-DNA sequencing identifies novel causative variants in *PSEN1*, *GRN* and *MAPT* in a clinical early-onset and familial Alzheimer's disease Ibero-American cohort. *Alzheimer's Res. Ther.* **4,** 34 (2012).
24. Benitez, B. A. *et al.* The *PSEN1*, p.E318G Variant Increases the Risk of Alzheimer's Disease in APOE-ε4 Carriers. *PLoS Genet.* **9,** e1003685 (2013).
25. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89,** 82–93 (2011).
26. Liang, W. S. *et al.* Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proc. Natl Acad. Sci. USA* **105,** 4441–4446 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Frohman for providing us with PLD1- and PLD2-WT constructs as well as constructs for the inactive mutations in these genes. This work was supported by grants from the National Institutes of Health (P30-NS069329, R01-AG044546 and R01-AG035083), the Alzheimer Association (NIRG-11-200110) and Barnes Jewish Foundation. This research was conducted while C.C. was a recipient of a New Investigator Award in Alzheimer's Disease from the American Federation for Aging Research. C.C. is a recipient of a BrightFocus Foundation Alzheimer's Disease Research Grant (A2013359S). Sequencing of some of the families included in this study was supported by Genentech and Pfizer. The recruitment and clinical characterization of research participants at Washington University were supported by NIH P50 AG05681, P01 AG03991 and P01 AG026276. This work was supported in part by the Intramural Research Program of the National Institute on Aging, National Institutes of Health, Department of Health and Human Services, project ZO1 AG000950-11. Samples from the National Cell Repository for Alzheimer's Disease (NCRAD) and NIA-LOAD, which receives government support under a cooperative agreement (U24 AG21886; U24: 5U24AG026395 and 1R01AG041797), were used in this study. We thank our contributors, including the Alzheimer's Disease Centers, that collected samples used in this study, as well as participants and their families, whose help and participation made this work possible. The Cache County Study is supported by National Institutes of Health, R01-AG11380, R01-AG18712 and R01-AG21136. Genotyping and analysis conducted at Brigham Young University was funded by grants from the National Institutes of Health R01-AG042611 and the Alzheimer's Association (MNIRG-11-205368) to J.S.K.K. The sequencing at University of Washington was supported by NIH R01039700 (Z.B.). The sequencing for the NIA-UK samples was supported by the Alzheimer's Research UK (ARUK), by an anonymous donor, by the NINDS (ZO1 AG000950-10), by the Wellcome Trust/MRC Joint Call in Neurodegeneration award (WT089698) to the UK Parkinson's Disease Consortium (UKPDC), by the Big Lottery (to K.M.) and by a fellowship from ARUK to R.G Some samples and pathological diagnoses were provided by the MRC London Neurodegenerative Diseases Brain Bank and the Manchester Brain Bank from Brains for Dementia Research, jointly funded from ARUK and AS via ABBUK Ltd. This work was also supported by the NIHR Queen Square Dementia BRU and BRC NIHR grant mechanisms. The sample recruitment and genetic studies at University of Pittsburgh are funded by NIH grants AG041718, AG030653, AG005133, AG07562 and AG023652. The Toronto sample studies are funded by Canadian Institutes of Health Research, Wellcome Trust, Medical Research Council, National Institute of Health, National Institute of Health Research, Ontario Research Fund and Alzheimer Society of Ontario (to P.S.G.-H.). The Nottingham Laboratory (K.M.) is funded by ARUK and Big Lottery. ARUK is supported by the UK Medical Research Council through the MRC Sudden Death Brain Bank (C.S.) and by a Project Grant (G0901254) and Training Fellowship (G0802462 to M.R.). P.P. receives funds from the Department of Health of the Government of Navarra, Spain (13085 and 3/2008) and from the UTE project FIMA, Spain. J.T.T. receives funds from the NIA (R01AG21136).

Author Contributions All the authors read and approved the manuscript. C.C. conceived and designed the experiments, supervised research, wrote the manuscript, performed the family and sample selection for exome-sequencing, and analysed the data. C.M.K., S.H., J.C. and A.T.J. performed all the cell-based analysis, and the PLD3 total gene-expression experiments. S.C.J. performed PLD3 pool-sequencing experiments. B.A.B. performed the genotyping of Val232Met and Ala442Ala in the Knight-ADRC and NIA-LOAD data sets, and analysed public gene-expression databases and carried out bioinformatic analysis of the effect of some variants on splicing. O.H., S.B. and Y.C. performed statistical and bioinformatic analyses. J.N. and D.L. recruited and assessed the NIA-LOAD families with the *PLD3* variants. J.B. T.S, D.C. and B.C. performed Sequenom genotyping. R.G., C.S., J.B., M.K.L., J.P., J.R.G., A.S., J.H. P.F., P.G.R., C.D.C., J.T.T., M.C.N., R.G.M., C.S., M.L., J.S.K.K., F.Y.D., M.N.B., X.W., O.L.L., M.G., M.I.K., C.M., J.T., J.L., A.B., I.B., K.B., K.M, O.L.B., P.P., Z.B., E.S., E.T., E.R. and P.S.G.-H., provided genotype data for the NIA-UK and NIMH datasets, Cache-County dataset, University of Pittsburgh dataset, University of Nottingham dataset, NIA-LOAD, the Wellderly dataset and the Toronto dataset. M.R. and D.G.H. performed the co-regulation pathway analysis. N.C. performed the neuropathological examination of the PLD3 Val232Met carriers. J.C.M. supervised recruitment and clinical assessment of the Knight-ADRC subjects, and A.M.G. supervised the functional and genetic experiments and critically reviewed all data and data analysis.

Author Information The authors declare competing financial interests: details are available in the online version of the paper. Exome-sequencing data is available on NIAGADs (https://www.niagads.org, accession number NG00033). Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.C. (ccruchaga@wustl.edu).

The UK Brain Expression Consortium members

John Hardy[3], Mina Ryten[3], Daniah Trabzuni[3], Michael E. Weale[30], Adaikalavan Ramasamy[30] & Colin Smith[31]

[30]Department of Medical and Molecular Genetics, King's College London, 16 De Crespigny Park, London SE5 8AF UK. [31]MRC Sudden Death Brain Bank Project, University of Edinburgh, South Bridge, Edinburgh EH8 9YL UK.

The ARUK Consortium members

Peter Passmore[32], David Craig[32], Janet Johnston[32], Bernadette McGuinness[32], Stephen Todd[32], Reinhard Heun[33], Heike Kölsch[34], Patrick G. Kehoe[35], Nigel M. Hooper[36], Emma R.L.C. Vardy[37], David M. Mann[38], Stuart Pickering-Brown[38], Kristelle Brown[18], Noor Kalsheker[18], James Lowe[18], Kevin Morgan[18], A. David Smith[39], Gordon Wilcock[39], Donald Warden[39] & Clive Holmes[40]

[32]Queen's University Belfast, University Road, Belfast BT7 1NN, UK. [33]Royal Derby Hospital, Uttoxeter Road, Derby, DE22 3NE, UK. [34]University of Bonn, Regina-Pacis-Weg 3, 53113 Bonn, Germany. [35]University of Bristol, Tyndall Avenue, Bristol, City of Bristol BS8 1TH, UK. [36]University of Leeds, Woodhouse Lane, Leeds, West Yorkshire LS2 9JT, UK. [37]University of Newcastle, Newcastle upon Tyne, Tyne and Wear NE1 7RU, UK. [38]University of Manchester, Oxford Road, Manchester, Greater Manchester M13 9PL, UK. [39]University of Oxford (OPTIMA), Wellington Square, Oxford OX1 2JD, UK. [40]University of Southampton, University Road, Southampton SO17 1BJ, UK.

METHODS

Participants and study design. The Institutional Review Board (IRB) at Washington University School of Medicine approved the study. Written informed consent was obtained from participants and their family members by the Clinical Core of the Knight ADRC. The approval number for the Knight ADRC Genetics Core is 93-0006.

Knight-ADRC samples. The Knight-ADRC sample included 1,114 late-onset Alzheimer's disease (LOAD) cases and 913 cognitively normal controls (377 older than 70 years), of European descent, and 302 African American Alzheimer's disease cases and controls, matched for age, gender and ethnicity. These individuals were evaluated by Clinical Core personnel of the Knight ADRC at Washington University. Cases received a clinical diagnosis of Alzheimer's disease dementia in accordance with standard criteria, dementia severity was determined using the Clinical Dementia Rating (CDR)[27].

Cerebrospinal fluid (CSF) levels data set: A subset ($n = 528$) of the Knight-ADRC samples had total tau protein and Aβ42 levels measured in the CSF by ELISA. Of these, 528, 303 did not have dementia (CDR = 0) and were elderly (over 65 years of age), with high CSF Aβ42 levels (>500 pg ml^{-1}). A description of the CSF data set used in this study can be found in another paper[11]. CSF collection and Aβ42, tau and phosphorylated tau181 measurements were performed as described previously[28].

NIA-LOAD. Participants from the National Institute of Ageing Late Onset Alzheimer Disease (NIA-LOAD) Family Study included a single individual with dementia from each of 868 families with at least three Alzheimer's disease-affected individuals, and 881 unrelated control individuals who were elderly and did not have dementia (545 individuals were older than 70 years of age). All Alzheimer's disease cases were diagnosed with dementia of the Alzheimer's type (DAT) using criteria equivalent to the National Institute of Neurological and Communication Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) for probable Alzheimer's disease[29]. NIA-LOAD families were ascertained based on the following criteria: probands (the affected individual through whom the family was recruited into the study) were required to have a diagnosis of definite or probable LOAD (onset after 60 years of age) and a sibling with definite, probable or possible LOAD with a similar age at onset. A third biologically related family member (first, second or third degree) was also required, regardless of affection status. This individual had to be ≥60 years of age if unaffected, or ≥50 years of age if diagnosed with LOAD or mild cognitive impairment[12]. Within each pedigree, we selected a single individual for the case–control series by identifying the youngest affected family member with the most definitive diagnosis (that is, individuals with autopsy confirmation were chosen over those with clinical diagnosis only). Unrelated controls without dementia who were used for the NIA-LOAD case–control series had no family history of Alzheimer's disease and were matched to the cases as previously described[12]. Only individuals of European descent based on the principal component (PC) factors from population stratification analyses were included. Written informed consent was obtained from all participants, and the study was approved by local IRB committees.

Wellderly Study. The Scripps Translational Science Institute's Wellderly study has recruited more than 1,000 healthy elderly participants. Inclusion criteria specify informed consent, age >80 years, blood or saliva donation, compliance with protocol-specified procedures, and no or mild ageing-related medical conditions. Exclusion criteria includes self-reported cancer (excluding basal and squamous cell skin cancer), coronary artery disease or myocardial infarction, stroke or transient ischaemic attack, deep vein thrombosis or pulmonary embolism, chronic renal failure or haemodialysis, Alzheimer's or Parkinson's disease, diabetes, aortic or cerebral aneurysm, or the use of oral chemotherapeutic agents, anti-platelet agents (excluding aspirin), cholinesterase inhibitors for Alzheimer's disease, or insulin. All genotyped individuals were of European descent.

Cache-County study. The Cache-County Study was initiated in 1994 to investigate the association of *APOE* genotype and environmental exposures on cognitive function and dementia. A cohort comprised of 5,092 Cache County, Utah, residents (representing 90% of all individuals in the county who were aged 65 or older) has been followed continually for over 15 years, completing four triennial waves of data collection including clinical assessments[13]. Genotypes were obtained for 255 demented individuals and 2,471 elderly cognitively normal individuals[13]. All individuals genotyped were of European descent.

UK-NIA data set. A description of the UK-NIA data set can be found in another paper[7]. In brief, this data set includes WES from 143 Alzheimer's disease cases and 183 elderly control individuals without dementia. All subjects were of European descent.

University of Pittsburgh data set. The *PLD3*(*V232M*) variant was genotyped in 2,211 subjects including 1,253 Alzheimer's disease cases (62.6% females) and 958 elderly control individuals without dementia(64.3% females). A complete description of the data set can be found in another paper[14]. All individuals were of European descent.

Toronto data set. The Toronto data set was composed of 269 unrelated Alzheimer's disease cases (53% females) and 250 unrelated controls without dementia (56% females) of European descent. The mean (s.d.) age at onset of Alzheimer's disease was 73 (±8) years, and the mean age (s.d.) at last examination of the controls was 73 (±10) years. The study was approved by the IRBs of the University of Toronto.

Exome sequencing. Enrichment of coding exons and flanking intronic regions was performed using a solution hybrid selection method with the SureSelect human all exon 50Mb kit (Agilent Technologies) following the manufacturer's standard protocol. This step was performed by the Genome Technology Access Center at Washington University. The captured DNA was sequenced by paired-end reads on the HiSeq 2000 sequencer (Illumina). Raw sequence reads were aligned to the reference genome hg19 using Novoalign (Novocraft Technologies). Base and SNP calling was performed by SNP Samtools. SNP annotation was carried out using version 5.07 of SeattleSeq Annotation server (see URL)[20].

On average, 95% of the exome had greater than eightfold coverage. SNP calls were made using SAM tools[30]. SNPs identified with a quality score lower than 20 and a depth of coverage lower than 5 were removed. More than 2,500 novel variants in the coding region were found per individual. We identified all variants shared by the affected individuals in a family. Variants not present in 1,000 genome project or the Exome Variant Server (EVS: http://evs.gs.washington.edu/EVS/) or with a frequency lower than 0.5% in the EVS were selected. On average, 80 coding variants were selected for each family. The selected variants were then genotyped in the remaining sampled family members. We validated more than 98% of the selected variants, confirming the high specificity of our exome-sequencing method and analysis. On average, we genotyped a total of 13 family members (7 cases and 6 controls) per family.

SNP genotyping. SNPs were genotyped using the Illumina Golden Gate, Sequenom, Kaspar and/or Taqman genotyping technologies. Only SNPs with a genotyping call rate higher than 98% and in Hardy–Weinberg equilibrium were used in the analyses. The principle of the MassARRAY system is PCR-based, with different size products analysed by SEQUENOM MALDI-TOF mass spectrometry[21,31]. The KBioscience Competitive Allele-Specific PCR (KASP) system is FRET-based endpoint-genotyping technology, v4.0 SNP (KBioscience)[21,31]. Genotype call rates were greater than 98%.

***PLD3* sequencing.** *PLD3* was sequenced in 2,363 cases and 2,027 controls of European origin, and 130 cases and 172 controls of African American descent using a pooled-DNA sequencing design as described previously[9,23,32]. In brief, equimolar amounts of individual DNA samples were pooled together following quantification using the Quant-iT PicoGreen reagent. Pools contained 100 ng of DNA per individual, from 94 individuals. The coding exons and flanking regions (a minimum of 50 bp each side) were individually PCR amplified using specific primers and Pfu Ultra high-fidelity polymerase (Stratagene). An average of 20 diploid genomes (approximately 0.14 ng DNA) per individual were used as input. PCR products were cleaned using QIAquick PCR purification kits, quantified using Quant-iT PicoGreen reagent and ligated in equimolar amounts using T4 Ligase and T4 Polynucleotide Kinase. After ligation, concatenated PCR products were randomly sheared by sonication and prepared for sequencing on an Illumina HighSeq2000 according to the manufacturer's specifications. pCMV6-XL5 amplicon (1,908 base pairs) was included in the reaction as a negative control. As positive controls, ten different constructs (*p53* gene) with synthetically engineered mutations at a relative frequency of one mutated copy per 188 normal copies was amplified and pooled with the PCR products.

Paired-end reads (101 bp) were aligned to the human genome reference assembly build 36.1 (hg19) using SPLINTER[32]. SPLINTER uses the positive control to estimate sensitivity and specificity for variant calling. The wild type: mutant ratio in the positive control is similar to the relative frequency expected for a single mutation in one pool (1 chromosome mutated in 94 samples = 1 in 188 chromosomes). SPLINTER uses the negative control (first 900 bp) to model the errors across the 101-bp Illumina reads and to create an error model from each sequencing run. Based on the error model SPLINTER calculates a *P* value for the probability that a predicted variant is a true positive. A *P* value at which all mutants in the positive controls were identified was defined as the cut-off value for the best sensitivity and specificity. All mutants included as part of the amplified positive control vector were found upon achieving >30-fold coverage at mutated sites (sensitivity = 100%) and only ~80 sites in the 1,908-bp negative control vector were predicted to be polymorphic (specificity = 95%). The variants with a *P* value below this cut-off value were considered for follow-up genotyping confirmation. All rare missense or splice-site variants were then validated by Sequenom and KASPar genotyping in each individual included in the pools. To avoid any batch or plate effects, cases and controls were included in each genotyping plate and all genotyping was performed in a single experiment. Finally, to confirm all of the

heterozygous calls, we created a custom DNA plate including all of the heterozygotes (cases and controls) for all of the variants, and then genotyped them again by Sequenom, creating a new Sequenom set.

Gene-expression and alternative splicing analyses. Total RNA was extracted using the RNeasy mini kit (Qiagen) following the manufacturer's protocol from 82 Alzheimer's disease cases and 39 individuals without dementia. Extracted RNA was treated with DNase1 to remove any potential DNA contamination. cDNAs were prepared from the total RNA, using the High-Capacity cDNA Archive kit (ABI). Gene-expression levels were analysed by real-time PCR, using an ABI-7900 real-time PCR system. The *PLD3*(*A442A*) variant was genotyped in DNA extracted from parietal lobe of 82 Alzheimer's disease cases and 39 individuals without dementia by KASPar as explained below. A total of eight carriers for the Ala442Ala variant were identified.

Total *PLD3* expression: gene expression was analysed by real-time PCR, using an ABI-7500 real-time PCR system. TaqMan assays were used to quantify *PLD3* mRNA levels. Primers and TaqMan probe for the reference gene, *GAPDH,* were designed over exon–exon boundaries, using Primer Express software, v3 (ABI) (sequences available on request). *Cyclophilin A* (ABI: 4326316E) was also used as a reference gene. Each real-time PCR run included within-plate triplicates and each experiment was performed at least twice for each sample.

Alternative splicing: we selected eight Ala442Ala carriers as well as eight CDR-, age-, *APOE*- and PMI-matched individuals to analyse the expression level of exon 11 containing transcripts, the exon in which the Ala442Ala variant is located. Real-time PCR assays were used to quantify *PLD3* exon 7 (forward primer, 5′-GCAGCTCCATCCCATCAACT-3′; reverse, 5′-CTTGGTTGTAGCGGGTGTCA-3′), exon 8 (forward primer, 5′-CTCAACGTGGTGGACAATGC-3′; reverse, 5′-AGTGGGCAGGTAGTTCATGACA-3′), 9 (forward primer, 5′-ACGAGCGTGGCGTCAAG-3′; reverse, 5′-CATGGATGGCTCCGAGTGT-3′), 10 (forward primer, 5′-GGTCCCCGCGGATGA-3′; reverse, 5′-GGTTGACACGGGCATATGG-3′) and 11 (first pair of primers: forward primer, 5′-CCAGCTGGAGGCCATTTTC-3′; reverse, 5′-TGTCAAGGTCATGGCTGTAAGG-3′; second pair forward primer, 5′-GCTGCTGGTGACGCAGAAT-3′; reverse, 5′-AGTCCCAGTCCCTCAGGAAAA-3′). Two pairs of primers were designed for exon 11 as an internal control. SYBR-green primers were designed using Primer Express software, v3 (ABI). Each real-time PCR run included within-plate duplicates and each experiment was performed at least twice for each sample. Real-time data were analysed using the comparative Ct method. Only samples with a standard error of <0.15% were analysed. The Ct values for exon 11 were normalized with the Ct value for the exons 7–10. The relative exon 11 levels for the Ala442Ala carriers versus the non-carriers were compared using a *t*-test.

***PLD3* gene expression in public databases.** We also used the GEO data sets GSE15222 (ref. 33) and GSE5281 (ref. 26) to analyse the association of *PLD3* gene expression and case-control status. In the GSE15222 data set, there are genotype and expression data from 486 late onset Alzheimer's Disease cases and 279 neuropathologically normal individuals without dementia. In the GSE5281 data set, samples were laser-captured from cortical regions of 16 normal elderly humans (10 males and 4 females) and from 33 Alzheimer's disease cases (15 males and 18 females). Mean age of cases and controls was 80 years. All samples were run on the Affymetrix U133 Plus 2.0 array. RNA data were re-normalized to an average expression of 8 units on a $\log_2$ scale. As potential covariates we analysed the brain region, gender and age for each sample. Stepwise discriminant analysis was used to identify the potential covariates to be included in the analysis of covariance (ANCOVA). For this data set we also extracted the gene-expression levels for *APP* (probe 211277_x_at), *PSEN1* (1559206_at) and *PSEN2* (203460_s_at) to examine the correlation between *PLD3* and *APP, PSEN1* and *PSEN2* using the Pearson correlation method.

Human brain samples and analysis of the Affymetrix Human Exon 1.0 ST array. Quantification and analysis of *PLD3* gene expression in brains was performed as previously described[34]. In brief, the human data used here were provided by the UK Human Brain Expression Consortium[34] and consisted of 101 control post-mortem brains. All samples originated from individuals with no significant neurological history or neuropathological abnormality and were collected by the MRC Edinburgh Brain Bank[35], ensuring a consistent dissection protocol and sample handling procedure. A summary of the available demographic details of these samples including a thorough analysis of their effects on array quality is provided in another paper[36]. All samples were accompanied by fully informed consent for retrieval and were authorized for ethically approved scientific investigation (Research Ethics Committee number 10/H0716/3). Total RNA was isolated from human post-mortem brain tissues using the miRNeasy 96-well kit (Qiagen). The quality of total RNA was evaluated by the 2100 Bioanalyzer (Agilent) and RNA 6000 Nano Kit (Agilent) before processing with the Ambion WT Expression Kit and Affymetrix GeneChip Whole Transcript Sense Target Labelling Assay and hybridization to the Affymetrix Exon 1.0 ST. All arrays were pre-processed using Robust Multi-array Average using Partek Genomics Suite v6.6 (Partek). The resulting expression data were corrected for individual effects (within which are nested post-mortem interval, brain pH, sex, age at death and cause of death) and experimental batch effects (date of hybridization). Transcript-level expression was calculated for 26,993 genes using Winsorized means (Winsorizing the data below 10% and above 90%).

RNA-pathway analysis. To evaluate the biological and functional relevance of co-expressed genes within the *PLD3*-containing modules, we used Weighted Gene Co-expression Network Analysis (WGCNA) and DAVID v6.7 (http://david.abcc.ncifcrf.gov/), the database for annotation, visualization and integrated discovery[37]. We restricted WGCNA to 15,409 transcripts that passed the Detection Above Background (DABG) criteria ($P < 0.001$ in at least 50% of samples in at least one brain region), had a coefficient of variation >5% and expression values exceeding 5 in all samples in at least one brain region. We followed a step-by-step network construction and module detection. In short, for each brain region, the Pearson correlations between all genes across all relevant samples were derived. We then calculated a signed-weighted co-expression adjacency matrix, allowing us to consider only positive correlations. A power 12, the default soft-threshold parameter for constructing a signed weighted network[38], was used in all brain regions, after checking that this threshold recapitulated scale-free topology[39]. Topological overlap, a more biologically meaningful measure of node interconnectedness (similarity)[9,23] than correlation, was subsequently calculated and genes were hierarchically clustered using 1 − topological overlap as the distance measure. Finally, modules were determined by using a dynamic tree-cutting algorithm. WGCNA led to the identification of several co-expression modules, ranging in number and size between the ten brain regions. We examined the overrepresentation (that is, enrichment) of the three Gene Ontology (GO) categories (biological processes, cellular components and molecular function) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways for each list of co-expressed genes with *PLD3* for each tissue by comparing numbers of significant genes annotated with this biological category with chance.

Statistical analyses. All of the single SNP analyses were performed using a Fisher's exact test, with no covariates included. Allelic association with risk for Alzheimer's disease was tested using 'proc logistic' in SAS including *APOE* genotype, age, PCs and study as covariates when available. Odds ratios with 95% confidence intervals and relative risks were calculated for the alternative allele compared to the most common allele using SAS. Association with age at onset (AAO) was carried out using the Kaplan–Meier method and tested for significant differences, using a proportional hazards model (proc PHREG, SAS) including gender and study as covariates. Controls without dementia were included in the analyses as censored data. The inclusion of these samples did not change the association. Gene-based analyses were performed using the optimal SNP-set (Sequence) Kernel Association Test (SKAT-O)[25].

Population attributable risk. We calculated the Population attributable risk (PAR) using the relative risk obtained in the study and the MAF from the EVS database (http://evs.gs.washington.edu/EVS/) and in the Cache-County data set, which is a population-based data set, using the equation:

$$\mathrm{PAR} = \frac{P_e(\mathrm{RR}_e - 1)}{(1 + P_e(\mathrm{RR}_e - 1))}$$

where P_e is the carrier frequency in the population and RR_e is the relative risk for the different variants.

Neuropathology studies. All study procedures were approved by Washington University's Human Research Protection Office. At autopsy, brain tissue was obtained from participants according to the protocol of the Knight-ADRC. Alzheimer's disease neuropathologic change was assessed according to the criteria of the National Institute on Ageing-Alzheimer's Association (NIA-AA)[40]. Dementia with Lewy bodies was assessed using the criteria given in another paper[41].

Cell-based studies. The following plasmids were used in this study: *pCMV6-XL5* human *PLD3-WT* (Origene), pCS2-Myc human *APP695-WT*[42], pCGN-*PLD-WT*[43] and Lys758Arg[44], pCGN-*PLD2-WT*[45] and Lys898Arg[44], pGFL-GFP[46], pGFP-V-RS-*PLD3*-shRNA-GI548821 (Origene) and pGFP-V-RS-Scr-shRNA-TR30013 (Origene). A dominant-negative mutation (Lys418Arg)[18] was introduced into the pCMV6-XL5 human *PLD3-WT* vector by site-directed mutagenesis using the QuikChangeII Site-Directed Mutagenesis kit (Agilent). All constructs were verified by Sanger sequencing.

Cell-culture assays. Human embryonic kidney (HEK293T) cells were cultured in Dulbecco's modified eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 1% L-glutamine and penicillin/streptomycin (solution containing penicillin and streptomycin). HEK293T cells were grown in 6-well lysine-coated plates. Mouse neuroblastoma (N2A) cells stably expressing human APP695 wild type were cultured in DMEM and Opti-MEM (50:50) supplemented with 5% FBS, 1% L -glutamine, penicillin/streptomycin and 500 μg ml^{-1} G418. After reaching

confluency, cells were transiently transfected with Lipofectamine 2000 (Invitrogen). Culture media were replaced after 24 h, and cells were incubated for another 24 h. Conditioned media were collected, treated with protease inhibitor cocktail and centrifuged at 3000*g* at 4 °C for 10 min to remove cell debris. Cell pellets were extracted on ice in lysis buffer (50 mM Tris, pH 7.6, 2 mM EDTA, 150 mM NaCl, 1% NP40, 0.5% Triton X-100, protease inhibitor cocktail) and centrifuged at 14,000*g*. Protein concentration was measured by the bicinchoninic acid (BCA) method as described by the manufacturer (Pierce-Thermo).

Real-time PCR and quantitative PCR. To confirm effective knockdown of endogenous mouse *PLD3* in mouse N2A-695 cells, RNA was extracted from cell lysates with an RNeasy kit (Qiagen) according to the manufacture's protocol. Extracted RNA (10 μg) was converted to cDNA by PCR using a High-Capacity cDNA Reverse Transcriptase kit (ABI). Gene expression was analysed by quantitative PCR (qPCR) using an ABI-7900 Real-Time PCR system (ABI). Taqman real-time PCR assays were used to quantify expression for mouse *PLD3* (Mm01171272_m1; ABI) and *GAPDH* (Hs02758991_g1; ABI). Samples were run in triplicate. To avoid amplication interference, expression asays were run in separate wells from the housekeeping gene *GAPDH*. Real-time data were analysed by the comparative C_T method. Average C_T values for each sample were normalized to the average C_T values for the housekeeping gene *GAPDH*. The resulting value was corrected for assay efficiency. Samples with a standard error of 20% or less were analysed.

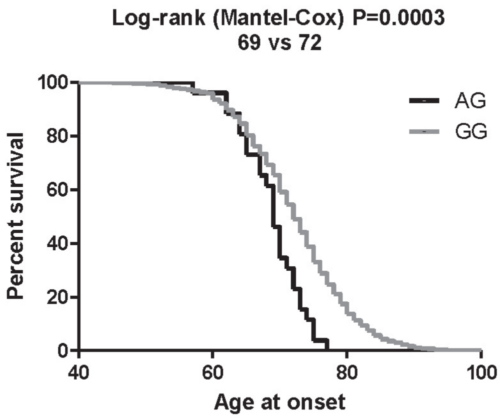
Immunoblot analysis. Standard SDS–PAGE was performed in 4–20% Criterion Tris-HCl gels (Bio-Rad). Samples were boiled for 5 min in Laemmli sample buffer before electrophoresis[47]. Immunoblots were probed with antibodies: PLD3 (Sigma), 9E10 (Sigma) and β-tubulin (Sigma).

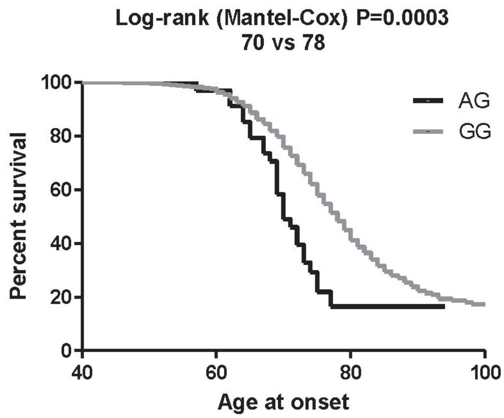
Enzyme-linked immunosorbent assay. The levels of Aβ40 and Aβ42 were measured in cell culture media by sandwich ELISA as described by the manufacturer (Invitrogen). ELISA values were obtained (measured in pg ml^{-1}) and corrected for total intracellular protein (measured in μg ml^{-1}) based on BCA measurements.

Immunoprecipitation. Cell lysates were incubated with Protein G beads (Thermo Scientific) to remove proteins from the solution that are prone to non-specifically bind to the beads (pre-cleared). Pre-cleared supernatants were incubated overnight at 4 °C with the antibodies indicated. Supernatant–antibody complexes were then incubated with Protein G beads at room temperature for 2 h. After washing, proteins were dissociated from the Protein G beads by incubating the beads in Laemmli sample buffer[47] supplemented with 5% β-mercaptoethanol at 95 °C for 10 min.

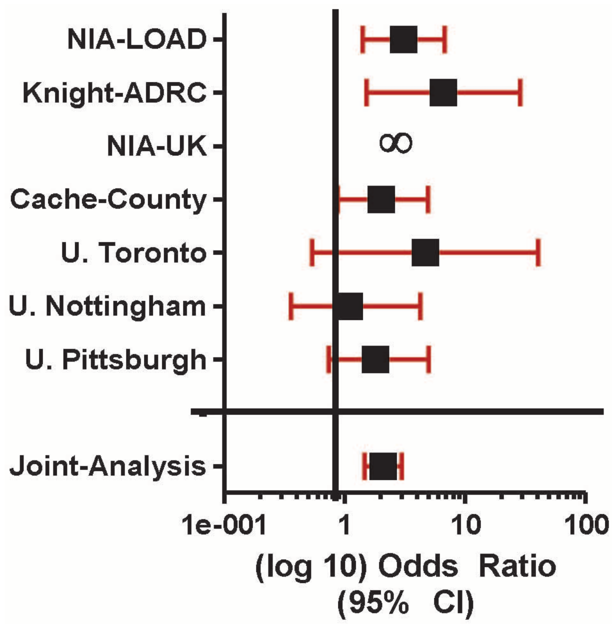
Bioinformatics analysis. SIFT (http://sift.jcvi.org/www/SIFT_BLink_submit.html) and Polyphen (http://genetics.bwh.harvard.edu/pph2/) algorithms were used to predict the functional effect of the identified variants. To determine the effect of the Ala442Ala variant on splicing we used the ESEfinder (http://rulai.cshl.edu/tools/ESE). Multiple sequence alignment was performed by ClustalW2, and the PLD3 orthologues were downloaded from Ensembl (http://www.ensembl.org/).

27. Morris, J. C. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* **43,** 2412–2414 (1993).
28. Fagan, A. M. *et al.* Inverse relation between *in vivo* amyloid imaging load and cerebrospinal fluid $A\beta_{42}$ in humans. *Ann. Neurol.* **59,** 512–519 (2006).
29. McKhann, G. *et al.* Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34,** 939–944 (1984).
30. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
31. Cruchaga, C. *et al.* SNPs associated with cerebrospinal fluid phospho-tau levels influence rate of decline in Alzheimer's disease. *PLoS Genet.* **6,** e1001101 (2010).
32. Vallania, F. L. *et al.* High-throughput discovery of rare insertions and deletions in large cohorts. *Genome Res.* **20,** 1711–1718 (2010).
33. Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nature Genet.* **39,** 1494–1499 (2007).
34. Forabosco, P., Ramasamy, A., Hardy, J. & Ryten, M. Insights into TREM2 biology by network analysis of human gene expression data. *Neurobiol. Aging* **34,** 2699–2714 (2013).
35. Millar, T. *et al.* Tissue and organ donation for research in forensic pathology: the MRC Sudden Death Brain and Tissue Bank. *J. Pathol.* **213,** 369–375 (2007).
36. Trabzuni, D. *et al.* Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *J. Neurochem.* **119,** 275–282 (2011).
37. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).
38. Mason, M. J., Fan, G., Plath, K., Zhou, Q. & Horvath, S. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics* **10,** 327 (2009).
39. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statist. Appl. Gen. Mol. Biol.* http:// dx.doi. org/10.2202/1544-6115.1128 (12 August 2005).
40. Montine, T. J. *et al.* National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. *Acta Neuropathol.* **123,** 1–11 (2012).
41. McKeith, I. G. *et al.* Diagnosis and management of dementia with Lewy bodies: third report of the DLB Consortium. *Neurology* **65,** 1863–1872 (2005).
42. Schroeter, E. H. *et al.* A presenilin dimer at the core of the γ-secretase enzyme: insights from parallel analysis of Notch 1 and APP proteolysis. *Proc. Natl Acad. Sci. USA* **100,** 13075–13080 (2003).
43. Hammond, S. M. *et al.* Characterization of two alternately spliced forms of phospholipase D1. *J. Biol. Chem.* **272,** 3860–3868 (1997).
44. Sung, T. C. *et al.* Mutagenesis of phospholipase D defines a superfamily including a *trans*-Golgi viral protein required for poxvirus pathogenicity. *EMBO J.* **16,** 4519–4530 (1997).
45. Colley, W. C. *et al.* Phospholipase D2, a distinct phospholipase D isoform with novel regulatory properties that provokes cytoskeletal reorganization. *Curr. Biol.* **7,** 191–201 (1997).
46. Kauwe, J. S. K. *et al.* Fine mapping of genetic variants in BIN1, CLU, CR1 and PICALM for association with cerebrospinal fluid biomarkers for Alzheimer's disease *PLoS One* **6,** e15918 (2011).
47. Cleveland, D. W., Fischer, S. G., Kirschner, M. W. & Laemmli, U. K. Peptide mapping by limited proteolysis in sodium dodecyl sulfate and analysis by gel electrophoresis. *J. Biol. Chem.* **252,** 1102–1106 (1977).

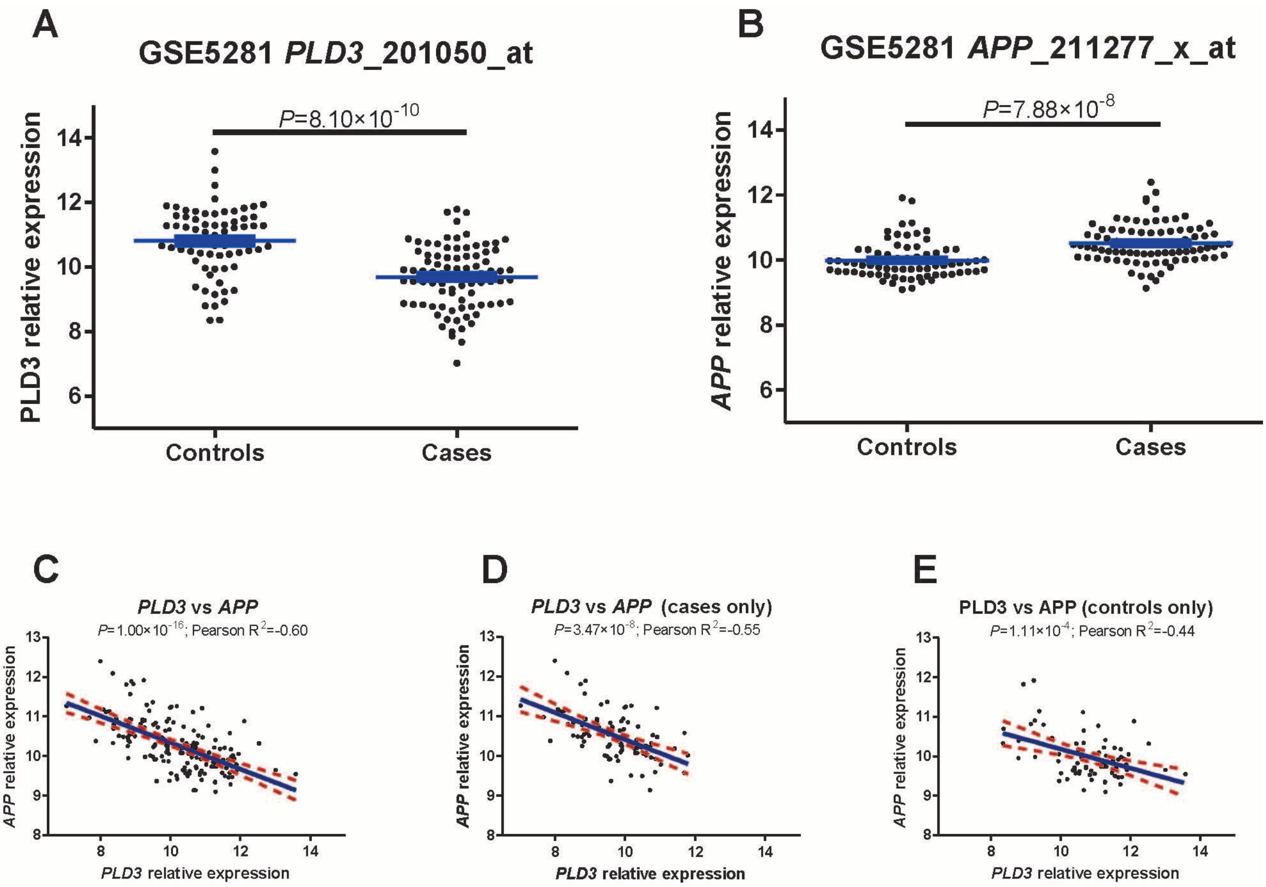




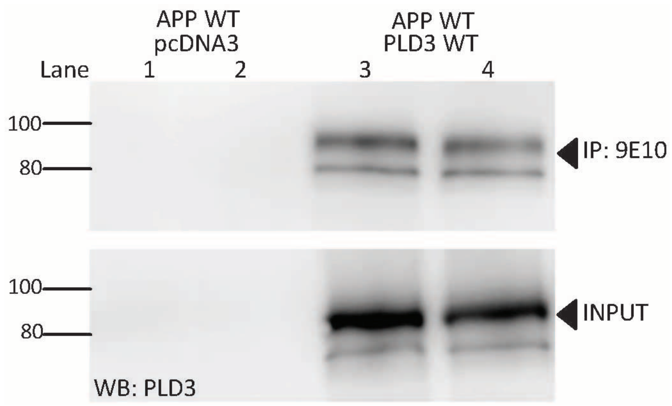
Extended Data Figure 1 | *PLD3*(*V232M*) is associated with age at onset for Alzheimer's disease. **a**, **b**, Age at onset was analysed for association with the *PLD3*(*V232M*) variant in 2,220 cases and 1,841 controls from the Knight-ADRC and NIA-LOAD data sets, by the Kaplan–Meier method. Data were tested for significant differences using the log-rank test. Case-only analysis (**a**); the carriers of the minor allele (AG) have an AAO 3 years lower than the non-carriers (69 versus 73; $P = 3 \times 10^{-3}$). Controls were included as censored data (**b**). The carriers of the minor allele have an AAO 8 years lower than the non-carriers (70 versus 78; $P = 3 \times 10^{-3}$). GG, homozygous for the GG genotype for the *PLD3*(*V232M*) variant.



Extended Data Figure 2 | Forest plot for each case–control series for the Val232Met variant.



Extended Data Figure 3 | *PLD3* and *APP* mRNA expression are inversely correlated. *PLD3* (probe 201050_at) and *APP* (probe 211277_x_at) expression levels were extracted from the GSE5281 data set. *PLD3* mRNA levels are significantly lower in Alzheimer's disease cases compared to controls ($P = 8.10 \times 10^{-10}$), but *APP* is higher in Alzheimer's disease cases ($P = 7.88 \times 10^{-8}$). *PLD3* mRNA levels are inversely correlated with *APP* mRNA expression levels ($P = 1.00 \times 10^{-16}$). The correlation is stronger in Alzheimer's disease cases (Person correlation coefficient = −0.55), than in controls (Person correlation coefficient = −0.44), but in both scenarios the correlation is highly significant.



Extended Data Figure 4 | PLD3 interacts with APP. HEK293T cells were transiently transfected with vectors containing *APP-WT* and an empty vector (pcDNA3) or *PLD3-WT* for 48 h. Cell lysates were extracted in non-ionic detergent, pre-cleared with Protein A beads and immunoprecipitated with an antibody to the Myc-tag on APP (9E10). Immunoblots were probed with an antibody specific to human PLD3. PLD1 and PLD2 reportedly do not inmunoprecipitate with APP[15,16].

Extended Data Table 1 | Association of the *PLD3*(*V232M*) variant in seven independent case–control data sets

Dataset	Cases	Carrier Freq %	Control	Carrier Freq %	OR (95%CI)	p-value
NIA-LOAD	29/1,077	2.62	8/920	0.86	3.09 (1.41-6.81)	4.00×10^{-03}
Knight-ADRC	16/1,098	1.44	2/911	0.22	6.63 (1.52-28.9)	3.40×10^{-03}
NIA-UK	1/142	0.70	0/183	0.00	∞ (NA)	0.438
Cache-County	6/249	2.35	29/2,442	1.17	2.03 (0.83-4.93)	0.131
U. Toronto	5/260	1.89	1/245	0.41	4.71 (0.54-40.7)	0.212
U. Nottingham	6/519	1.14	3/271	1.09	1.05 (0.26-4.25)	1.000
U. Pittsburgh	15/1,253	1.18	6/958	0.62	1.82 (0.74-5.00)*	0.191
NIMH	4/318	1.24	-	-	N/A	
Wellderly	-	-	1/376	0.27	N/A	
Total	**82/4,916**	**1.64**	**50/6,306**	**0.79**	**2.10 (1.47-2.99)**	$\mathbf{2.93\times10^{-05}}$

The table shows the counts for carriers and non-carriers. *P* values were calculated by Fisher's exact-test. *For the University of Pittsburgh data set, age, gender, *APOE* genotype and principal component factors for population stratification were available. Association of the Val232Met variant with Alzheimer's disease risk was performed by logistic regression including age, sex, *APOE* genotype and the first four principal component factors as covariates. N/A, not applicable.

Extended Data Table 2 | Sequence variants found in *PLD3* in the NIA-LOAD, Knight-ADRC and NIA-UK data sets

Chr. position	AA		NIA LOAD	Knight ADRC	NIA-UK	total	MAF %	p-value	OR (95% CI)	EVS MAF%	SIFT	Polyphen
40872407	M6R	CA	0	8	1	9	0.19	0.02	7.73	NP	tolerated	deleterious
		CO	0	1	0	1	0.02		(1.09-61)			
40872764	S63G	CA	3	1	0	4	0.08	0.74	0.68	0.16	tolerated	neutral
		CO	5	0	0	5	0.12		(0.18-2.55)			
40872803	P76A	CA	3	1	0	4	0.08	0.12	NA	0.03	tolerated	benign
		CO	0	0	0	0	0.00					
40873764	T136M	CA	0	1	0	1	0.02	0.54	NA	NP	tolerated	deleterious
		CO	0	0	0	0	0.00					
40876055	H197Y	CA	0	1	0	1	0.02	0.49	0.85	NP	damaging	benign
		CO	0	1	0	1	0.02		(0.05-13-7)			
40877584	K228R	CA	1	1	1	3	0.06	0.25	NA	NP	damaging	deleterious
		CO	0	0	0	0	0.00					
40877595	V232M	CA	29	16	1	46	0.99	**$1.05x10^{-05}$**	3.99	0.48	damaging	deleterious
		CO	8	2	0	10	0.25		(2.01-7.94)			
40877608	N236S	CA	0	2	0	2	0.04	0.40	1.71	0.01	damaging	deleterious
		CO	0	1	0	1	0.02		(0.15-18.91)			
40877752	N284S	CA	0	1	0	1	0.02	0.54	NA	NP	tolerated	deleterious
		CO	0	0	0	0	0.00					
40880407	C300Y	CA	2	3	0	5	0.10	0.46	2.14	0.09	tolerated	deleterious
		CO	1	0	1	2	0.04		(0.41-11.06)			
40880481	A325T	CA	0	1	0	1	0.02	0.54	NA	NP	damaging	deleterious
		CO	0	0	0	0	0.00					
40883725	Q406H	CA	1	0	0	1	0.02	0.54	NA	NP	tolerated	neutral
		CO	0	0	0	0	0.00					
40883783	T426A	CA	1	0	0	1	0.02	0.54	NA	NP	tolerated	neutral
		CO	0	0	0	0	0.00					
40883911	G435V	CA	0	0	0	0	0.00	0.46	NA	0.02	damaging	deleterious
		CO	1	0	0	1	0.02					
40883933	A442A	CA	48	35	12	95	2.09	**$1.08x10^{-05}$**	2.31	1.59	-	-
		CO	17	12	7	36	0.90		(1.56- 3.41)			
40883956	Q450L	CA	0	0	0	0	0.00	0.46	NA	NP	tolerated	neutral
		CO	0	0	1	1	0.02					
40883962	G452E	CA	4	6	0	10	0.21	0.16	2.86	0.09	tolerated	deleterious
		CO	0	2	1	3	0.07		(0.78-10.4)			
40883967	G454C	CA	0	1	0	1	0.02	0.54	NA	NP	damaging	deleterious
		CO	0	0	0	0	0.00					
40884037	D477G	CA	0	1	0	1	0.02	0.49	0.42	0.02	damaging	deleterious
		CO	0	1	0	1	0.02		(0.04- 4.72)			
40884069	R488C	CA	0	3	0	3	0.06	0.25	NA	0.02	damaging	deleterious
		CO	0	0	0	0	0.00					
total		CA	1106	1114	143	2363						
total		CO	928	913	183	2024						

The coding region of *PLD3* was sequenced in 2,363 Alzheimer's disease cases and 2,024 controls (see Methods) from the Knight-ADRC, NIA-LOAD and the NIA-UK data sets. The table shows the coding variants identified as well as the number of carriers in each data set. The minor allele frequency (MAF) in cases and in controls, the *P* value and the odds ratio (OR) for the association with case-control status is shown. The MAF of the identified variants in the Exome Variant Server (EVS) is shown. We also used SIFT and Polyphen to predict the impact of the non-synonymous changes on protein function. AA, amino acid; CA, cases; CO, controls; NA, not applicable; NP, not present.

Extended Data Table 3 | Gene-based analysis including all coding variants or only variants predicted to be deleterious

	Benign + deleterious		Only deleterious	
	p-value	OR (CI)	p-value	OR (CI)
All variants	1.44×10^{-11}	2.75 (2.05-3.68)	2.52×10^{-12}	2.86 (2.10-3.88)
Excluding V232M	1.58×10^{-8}	2.58 (1.87-3.57)	2.95×10^{-8}	2.54 (1.81-3.57)
Excluding A442 and V232M	1.61×10^{-3}	2.86 (1.62-5.06)	5.88×10^{-5}	3.20 (1.59-6.45)

Gene-based analyses were performed using SKAT-O. Variants that were predicted to be benign by both SIFT and Polyphen were removed for the second analysis.

Extended Data Table 4 | Association analysis for *PLD3*(*A442A*) in four data sets of individuals of European descent

	CA	CO	p-value	OR (95% CI)
NIA-LOAD	48/1058	17/911	1.40×10^{-03}	2.43 (1.38-4.25)
Knight-ADRC	35/1079	12/901	7.10×10^{-03}	2.43 (1.25-4.71)
NIA-UK	12/131	7/176	9.76×10^{-02}	2.30 (0.88- 6.0)
Cache-County	9/246	50/2421	1.15×10^{-01}	1.77 (0.86-3.65)
Total	104/2514	86/4409	3.78×10^{-07}	2.12 (1.58-2.83)

The table shows the counts for carriers and non-carriers. *P* values were calculated using the Fisher's exact test. CA, cases; Co, controls.

Extended Data Table 5 | *PLD3* is associated with risk for Alzheimer's disease in African Americans

	Cases (n=130)		Controls (n=172)			
Variant	carriers	Carrier Freq %	carriers	Carrier Freq %	p-value	OR (95% CI)
G63S	1	0.77%	0	0.00%	0.43	NA
K228R	1	0.77%	0	0.00%	0.43	NA
V232M	3	2.31%	0	0.00%	**0.07**	NA
I364I	6	4.62%	4	2.33%	0.33	2.02 (0.56-7.29)
A442A	4	3.08%	0	0.00%	**0.03**	NA
Total	15	11.54%	4	2.33%	**1.4×10^{-03}**	5.48 (1.77-16.92)

PLD3 was sequenced in a total of 302 African Americans. The table shows the counts for single SNPs and the gene-based analysis for *PLD3* in 130 African American cases and 172 controls. *P* values were calculated using the Fisher's exact test. NA, not applicable.

LETTER

doi:10.1038/nature12932

Oestrogen increases haematopoietic stem-cell self-renewal in females and during pregnancy

Daisuke Nakada[1,2,3], Hideyuki Oguro[4], Boaz P. Levi[5], Nicole Ryan[1]‡, Ayumi Kitano[1], Yusuke Saitoh[1], Makiko Takeichi[1], George R. Wendt[5] & Sean J. Morrison[4]

Sexually dimorphic mammalian tissues, including sexual organs and the brain, contain stem cells that are directly or indirectly regulated by sex hormones[1–6]. An important question is whether stem cells also exhibit sex differences in physiological function and hormonal regulation in tissues that do not show sex-specific morphological differences. The terminal differentiation and function of some haematopoietic cells are regulated by sex hormones[7–10], but haematopoietic stem-cell function is thought to be similar in both sexes. Here we show that mouse haematopoietic stem cells exhibit sex differences in cell-cycle regulation by oestrogen. Haematopoietic stem cells in female mice divide significantly more frequently than in male mice. This difference depends on the ovaries but not the testes. Administration of oestradiol, a hormone produced mainly in the ovaries, increased haematopoietic stem-cell division in males and females. Oestrogen levels increased during pregnancy, increasing haematopoietic stem-cell division, haematopoietic stem-cell frequency, cellularity, and erythropoiesis in the spleen. Haematopoietic stem cells expressed high levels of oestrogen receptor-α (ERα). Conditional deletion of ERα from haematopoietic stem cells reduced haematopoietic stem-cell division in female, but not male, mice and attenuated the increases in haematopoietic stem-cell division, haematopoietic stem-cell frequency, and erythropoiesis during pregnancy. Oestrogen/ERα signalling promotes haematopoietic stem-cell self-renewal, expanding splenic haematopoietic stem cells and erythropoiesis during pregnancy.

A fundamental question in stem-cell biology concerns the extent to which stem cells are regulated by long-range signals to ensure that stem-cell function within individual tissues is integrated with the overall physiological state[11]. For example, stem cells in the intestine, central nervous system and germ line are regulated by insulin and nutritional status[12–17]. Among haematopoietic cells oestrogen regulates proliferation, survival, differentiation and cytokine production by lymphoid and myeloid cells[10,18,19], and induces apoptosis in erythroid cells by inhibiting GATA1 (refs 20, 21). This raises the question of whether sex hormones also regulate haematopoietic stem cells.

Comparing 8–10-week-old male and female mice, we observed no significant differences in the frequency (Fig. 1a) or total numbers (Fig. 1b, c) of $CD150^{+}CD48^{-}Lin^{-}Sca\text{-}1^{+}c\text{-}kit^{+}$ haematopoietic stem cells or $CD150^{-}CD48^{-}Lin^{-}Sca\text{-}1^{+}c\text{-}kit^{+}$ multipotent progenitors (MPPs)[22], or in the percentage of bone marrow cells that incorporated a 10-day pulse of 5-bromodeoxyuridine (BrdU; Fig. 1d). However, a significantly higher percentage of haematopoietic stem cells and MPPs incorporated BrdU in female as compared to male mice (Fig. 1d). Because the haematopoietic stem cells had incorporated BrdU while remaining in the haematopoietic stem-cell pool, haematopoietic stem cells undergo more frequent self-renewing divisions in female mice than in male mice.

To test this using an independent approach we treated 4–6-week-old *Rosa26-rtTA; tetO-H2B-GFP* mice[23] with doxycycline for 6 weeks to induce histone H2B–GFP expression and then chased for 12 weeks without doxycycline to assess the rate of H2B–GFP dilution as a result of cell division. After 6 weeks of doxycycline treatment, haematopoietic stem cells, MPPs and whole bone marrow (WBM) cells in male and female mice were strongly and uniformly labelled with H2B–GFP (Fig. 1e). However, after the 12-week chase almost all bone marrow cells lost H2B–GFP expression in male and female mice (Fig. 1e, f). As expected[23,24], haematopoietic stem cells and MPPs retained substantial frequencies of H2B–GFP^{high} cells that were relatively quiescent during the chase period (Fig. 1e, f). Consistent with the higher rate of BrdU incorporation in female haematopoietic stem cells, significantly ($P <$ 0.005) lower percentages of haematopoietic stem cells and MPPs retained high levels of H2B–GFP in female as compared to male mice (Fig. 1e, f). Haematopoietic stem cells and MPPs thus divide more frequently in female as compared to male mice.

Ovariectomy, but not castration, significantly reduced the percentage of haematopoietic stem cells and MPPs that incorporated a 10-day pulse of BrdU (Fig. 2a). Indeed, ovariectomy reduced haematopoietic stem-cell and MPP division in females to male levels (Fig. 2a). Castration or ovariectomy did not affect the numbers of haematopoietic stem cells or MPPs in the bone marrow (Extended Data Fig. 1a) and produced only minor changes in the gross lineage composition of bone marrow cells (Extended Data Fig. 1b). The rate of haematopoietic stem-cell division in female mice is therefore increased by signals from the ovary.

To test whether female sex hormones can affect haematopoietic stem-cell cycling we administered oestradiol (E2; $2\,\mu g\,day^{-1}$), progesterone (P; $1\,mg\,day^{-1}$)[5], or oestradiol with progesterone (E2+P) to young adult male and female mice for 1 week along with BrdU for the last 3 days. This significantly increased oestrogen and/or progesterone levels in both male and female mice (Extended Data Fig. 3a, b) without exceeding the physiological levels observed during pregnancy. These treatments did not affect bone marrow or spleen cellularity (Fig. 2b) or haematopoietic stem-cell frequency (Fig. 2c), but E2 induced erythropoiesis in the spleen (Extended Data Fig. 2d). Treatment with E2 or E2+P, but not P alone, significantly increased BrdU incorporation by haematopoietic stem cells, but not by unfractionated bone-marrow cells, in both male and female mice (Fig. 2d).

E2 treatment increased BrdU incorporation by haematopoietic stem cells in castrated and ovariectomized mice, indicating that E2 acts independently of the gonads (Fig. 2e). E2 treatment also increased the frequency of $Ki67^{+}$ haematopoietic stem cells (Fig. 2f and Extended Data Fig. 4a). E2 treatment significantly reduced the frequency of haematopoietic stem cells that retained H2B–GFP in *Rosa26-rtTA; tetO-H2B-GFP* mice (Extended Data Fig. 4b, c). In contrast, treatment with dihydrotestosterone did not affect BrdU incorporation by male or female haematopoietic stem cells (Fig. 2g) or haematopoietic stem-cell frequency in bone marrow (Fig. 2h).

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. [2]Stem Cells and Regenerative Medicine Center, Baylor College of Medicine, Houston, Texas 77030, USA. [3]Center for Cell and Gene Therapy, Baylor College of Medicine, Houston, Texas 77030, USA. [4]Howard Hughes Medical Institute, Department of Pediatrics, and Children's Research Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [5]Life Sciences Institute, Center for Stem Cell Biology, University of Michigan, Ann Arbor, Michigan 48109, USA.
‡Deceased.

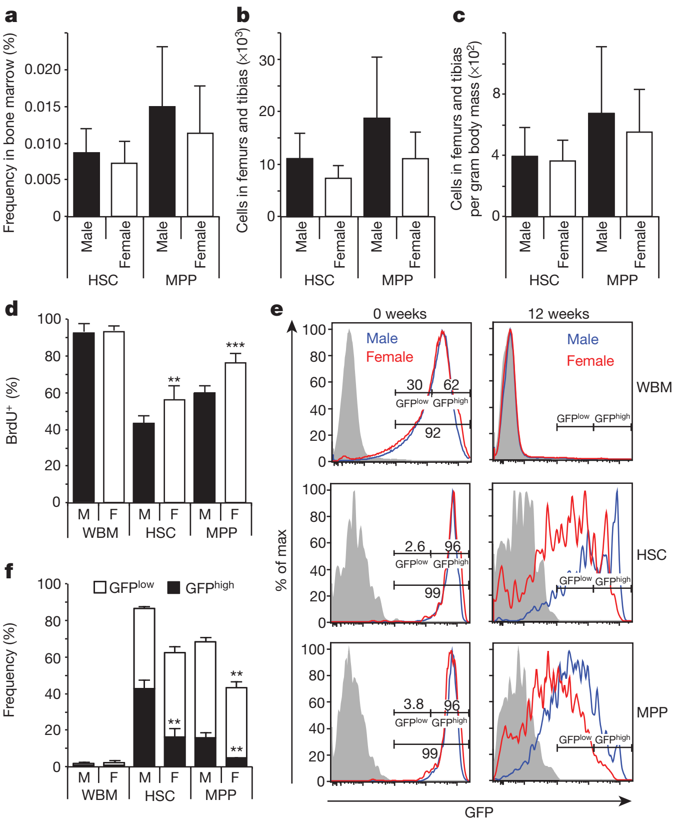


Figure 1 | Haematopoietic stem cells divide more frequently in female mice than in male mice. a–c, The frequency of haematopoietic stem cells (HSCs) and MPPs in the bone marrow (**a**), the total numbers of haematopoietic stem cells and MPPs in two femurs and tibias (**b**), and the numbers of haematopoietic stem cells and MPPs per gram of body mass (**c**) did not differ between young adult male and female mice. **d**, BrdU incorporation into whole bone marrow (WBM) cells, haematopoietic stem cells and MPPs during a 10-day pulse (**a–d**, $n = 5$ mice per group in five independent experiments). **e**, H2B–GFP intensity immediately after a 6-week pulse of doxycycline in *Col1A1-H2B-GFP; Rosa26-M2-rtTA* mice (left) or after a 12-week chase without doxycycline (right). **f**, The percentages of WBM cells, haematopoietic stem cells and MPPs that retained H2B–GFP (4 males and 3 females in 3 independent experiments). All data represent mean ± standard deviation; $*P < 0.05$; $**P < 0.005$; $***P < 0.0005$ by Student's *t*-test.

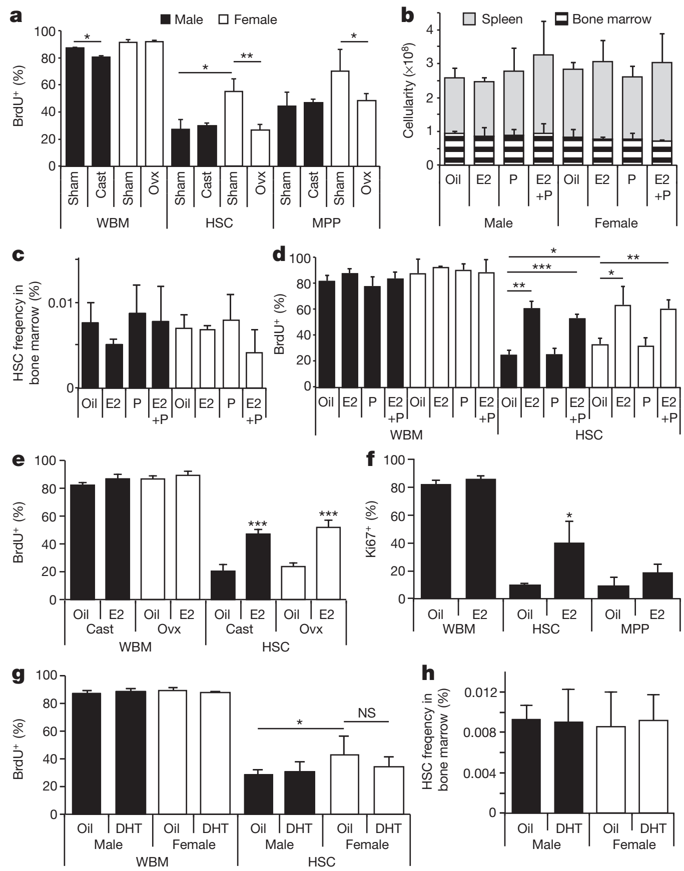


Figure 2 | Increased haematopoietic stem-cell division in female mice depends on the ovaries and is stimulated by oestradiol. a, Effect of castration (cast) or ovariectomy (ovx) on the rates of division by WBM cells, haematopoietic stem cells, or MPPs (3 sham and 4 gonadectomized mice in 3 independent experiments). **b**, **c**, Administering oestradiol (E2), progesterone (P), or both (E2+P) for 1 week did not affect the number of bone marrow cells or splenocytes (**b**), or haematopoietic stem-cell frequency in bone marrow (**c**). Oil, corn oil. **d**, Administering E2 or E2+P significantly increased haematopoietic stem-cell division in male and female mice (**b–d**, $n = 3$ mice per treatment in three independent experiments). **e**, Administering E2 to castrated or ovariectomized mice significantly increased haematopoietic stem-cell division by BrdU incorporation ($n = 5$). **f**, Administering E2 to male mice increased the frequency of haematopoietic stem cells positive for Ki67 ($n = 3$). **g**, **h**, Administering dihydrotestosterone (DHT) for 7 days did not affect haematopoietic stem-cell division or haematopoietic stem-cell frequency ($n = 4$ mice per treatment in 4 independent experiments). Data represent mean ± standard deviation; $*P < 0.05$; $**P < 0.005$; $***P < 0.0005$ by Student's *t*-test. NS, not significant.

Consistent with the observation that oestrogen induces apoptosis in erythroid progenitors[20], we observed an increased frequency of annexin-V$^+$Ter119$^+$ cells in female as compared to male bone marrow (Extended Data Fig. 5a). This appeared to be offset by increased generation of megakaryocyte–erythroid progenitors (MEPs) in female mice (Extended Data Fig. 5b). Neither MEPs nor Ter119$^+$ cells exhibited differences in cell-cycle distribution between males and females (Extended Data Fig. 5c). Given that MEPs may arise directly from the asymmetric division of HSCs[25], these observations raise the possibility that the increased frequency of MEPs in female mice reflects increased asymmetric self-renewal of female haematopoietic stem cells in response to oestrogen.

We treated mice for 14 days with the aromatase inhibitor anastrozole, which reduces oestrogen levels[26]. Anastrozole did not significantly affect bone marrow cellularity (Fig. 3a) or lineage composition (Extended Data Fig. 6a), but slightly reduced haematopoietic stem-cell frequency in female mice (Fig. 3b). Anastrozole did not significantly affect BrdU incorporation (during the last 10 days of anastrozole) by whole bone marrow cells or MPPs in male or female mice, but did significantly reduce BrdU incorporation by female haematopoietic stem cells ($P < 0.05$, Fig. 3c). Treatment with the progesterone receptor antagonist RU486 had no effect on bone marrow or spleen cellularity, haematopoietic stem-cell frequency, or BrdU incorporation (Extended Data Fig. 3d–f). These results indicated that endogenous oestrogen increases haematopoietic stem-cell division in female mice.

Haematopoietic stem cells and MPPs from male and female mice expressed high levels of oestrogen receptor-α (ERα; encoded by *Esr1*) (Fig. 3d, e). However, haematopoietic stem cells expressed little or no ERβ (encoded by *Esr2*), progesterone receptor (*Pgr*), or androgen receptor (*Ar*) (Fig. 3d). To assess the roles of ERα and ERβ in haematopoietic stem-cell regulation we treated male mice with the ERα agonist propylpyrazoletriol (PPT) or the ERβ agonist diarylpropionitrile (DPN)[27] for 2 weeks along with BrdU for the last 10 days. PPT and DPN did not affect bone marrow or spleen cellularity, or the frequencies of haematopoietic stem cells and MPPs (Extended Data Fig. 7a, b). PPT, but not DPN, significantly increased erythropoiesis in the bone marrow and spleen (Extended Data Fig. 7c) as well as BrdU incorporation by haematopoietic stem cells (Extended Data Fig. 7d), indicating that oestrogen effects on haematopoietic stem cells are mediated mainly by ERα. Consistent with this conclusion, germline *Esr1*-deficient mice of both sexes had normal bone marrow cellularity and lineage composition

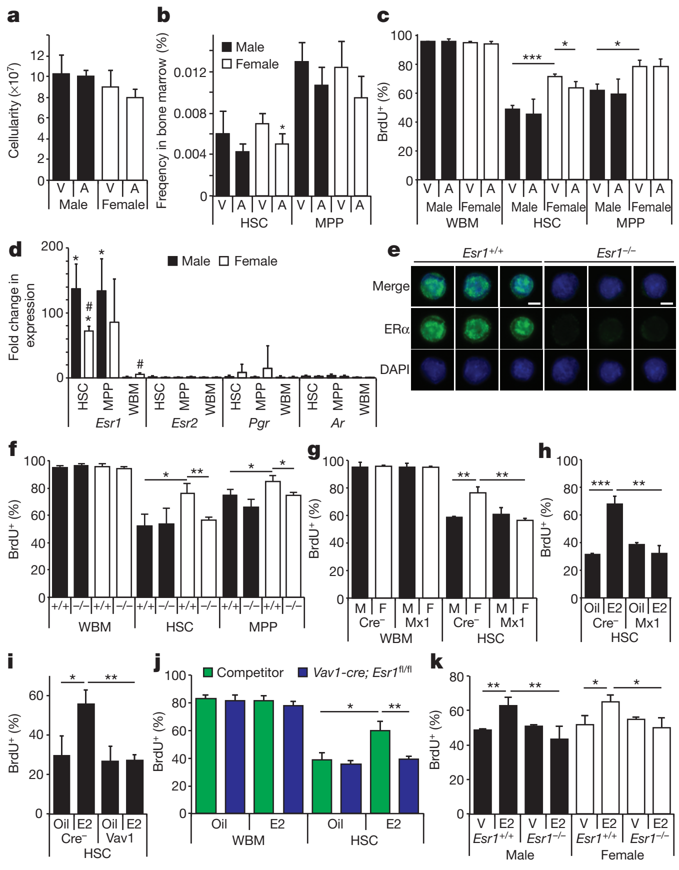


Figure 3 | Oestradiol–ERα signalling promotes haematopoietic stem-cell division in female mice. **a**, **b**, Bone marrow cellularity (**a**) and haematopoietic stem-cell and MPP frequency (**b**) in mice administered the aromatase inhibitor anastrozole (A) or vehicle (PBS, abbreviated V) for 2 weeks. **c**, BrdU incorporation (10-day pulse) by WBM cells, haematopoietic stem cells, or MPPs in male or female mice treated with anastrozole or vehicle (**a**–**c**, 4 PBS-treated and 6 anastrozole-treated mice in 4 independent experiments). **d**, qRT–PCR revealed that haematopoietic stem cells and MPPs from female and male mice expressed greatly elevated levels of *Esr1* (which encodes ERα) but not *Esr2* (ERβ), *Pgr* (progesterone receptor), or *Ar* (androgen receptor) relative to male WBM (*$P < 0.05$ between HSC/MPP and WBM; #$P < 0.05$ between male and female). Expression levels were normalized to β-actin. **e**, Immunofluorescence for ERα in haematopoietic stem cells (**d**, **e**, $n = 3$ mice from 3 experiments). Scale bar, 4 μm. **f**, BrdU incorporation (10-day pulse) by WBM cells, haematopoietic stem cells and MPPs in male and female mice (−/−, *Esr1*-deficient; +/+, littermate controls, **f**–**h**, $n = 3$ mice per group in 3 independent experiments). **g**, Conditional deletion of *Esr1* in female *Mx1-cre; Esr1*$^{fl/fl}$ mice reduced BrdU incorporation into haematopoietic stem cells (Cre$^-$, *Esr1*$^{fl/fl}$; Mx1, *Mx1-cre; Esr1*$^{fl/fl}$, $n = 3$). **h**, **i**, Conditional deletion of *Esr1* in male *Mx1-cre; Esr1*$^{fl/fl}$ mice (**h**) or *Vav1-cre; Esr1*$^{fl/fl}$ mice (**i**) rendered haematopoietic stem cells insensitive to exogenous oestrogen (**h**, **i**, $n = 3$ mice per group in 2 independent experiments). **j**, E2 treatment of mice competitively reconstituted with WBM cells from wild-type and *Vav1-cre; Esr1*$^{fl/fl}$ mice increased BrdU incorporation by wild-type haematopoietic stem cells but not *Esr1*-deficient haematopoietic stem cells (3 oil-treated and 4 E2-treated mice in 2 independent experiments). **k**, Effect of E2 on haematopoietic stem cells freshly added to culture (serum-free, phenol-red-free medium with E2 or vehicle for 3 days; BrdU for 1 h; $n = 3$ mice in 2 independent experiments). All data represent mean ± standard deviation; * and #, $P < 0.05$; **$P < 0.005$; ***$P < 0.0005$ by Student's *t*-test.

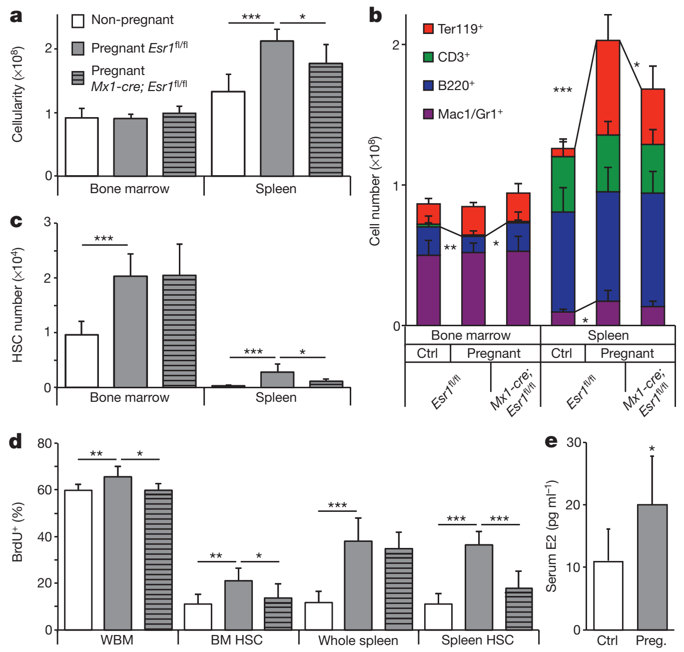


Figure 4 | Increased haematopoietic stem-cell division, haematopoietic stem-cell frequency, and erythropoiesis in the spleen during pregnancy depend on ERα signalling in haematopoietic cells. **a**, Spleen and bone marrow cellularity. Pregnant mice were on day 14.5 of gestation. **b**, Pregnant mice had significantly increased Mac1/Gr1$^+$ myeloid cells, Ter119$^+$ erythroid cells, and overall cellularity in the spleen, but reduced bone marrow B220$^+$ B cells. The increase in splenic erythropoiesis required ERα expression by haematopoietic cells. **c**, Haematopoietic stem-cell frequency in the bone marrow and spleen. **d**, In pregnant mice the rate of BrdU incorporation (24-h pulse) significantly increased in whole bone marrow (WBM) cells, bone marrow haematopoietic stem cells, and spleen haematopoietic stem cells and depended on ERα expression by haematopoietic cells (**a**–**d**, 9 non-pregnant, 7 pregnant *Esr1*$^{fl/fl}$, and 6 pregnant *Mx1-cre; Esr1*$^{fl/fl}$ mice in 9 independent experiments). **e**, Serum E2 levels in mice (21 non-pregnant and 9 pregnant mice from 6 independent experiments). All data represent mean ± standard deviation; *$P < 0.05$; **$P < 0.005$; ***$P < 0.0005$ by Student's *t*-test.

(Extended Data Fig. 6b, d), as well as normal haematopoietic stem-cell and MPP frequency (Extended Data Fig. 6c), but significantly reduced BrdU incorporation into haematopoietic stem cells in female but not male mice (Fig. 3f).

To test whether ERα acts autonomously in haematopoietic stem cells we conditionally deleted *Esr1* from haematopoietic cells. *Mx1-cre; Esr1*$^{fl/fl}$ mice and *Esr1*$^{fl/fl}$ controls were treated with polyinosinic:polycytidylic acid (poly(I:C); four doses of 10 μg per 20 g body mass per day over 8 days) to induce *Mx1-cre* expression, then 19–21 days after poly(I:C) treatment we pulsed with BrdU for 10 days. Conditional deletion of *Esr1* from haematopoietic cells significantly reduced BrdU incorporation into haematopoietic stem cells in female, but not male, mice (Fig. 3g).

Seven days of E2 significantly increased BrdU incorporation (3-day pulse) by haematopoietic stem cells from *Esr1*$^{fl/fl}$ controls but not *Mx1-cre; Esr1*$^{fl/fl}$ mice (Fig. 3h). Similar results were obtained using *Vav1-cre; Esr1*$^{fl/fl}$ mice (Fig. 3i), indicating that *Esr1*-deficient haematopoietic stem cells are not capable of responding to exogenous oestrogen.

To test whether E2 acts directly on haematopoietic stem cells, we competitively transplanted 10^6 CD45.2$^+$ *Vav1-cre; Esr1*$^{fl/fl}$ bone marrow cells along with 10^6 CD45.1$^+$ bone marrow cells into irradiated mice. Fifteen weeks later we treated the mice with either E2 or vehicle for 7 days along with BrdU for the last 3 days. E2 treatment did not significantly affect BrdU incorporation by wild-type or *Esr1*-deficient bone marrow cells (Fig. 3j). E2 treatment did significantly increase BrdU incorporation by wild-type haematopoietic stem cells but not by *Esr1*-deficient haematopoietic stem cells in the same mice (Fig. 3j). This demonstrates that E2 acts directly on haematopoietic stem cells, rather than acting indirectly by stimulating secondary signals from other haematopoietic stem cells. Consistent with this, addition of E2 to cultured haematopoietic stem cells significantly increased BrdU incorporation by wild-type haematopoietic stem cells from male and female mice but not *Esr1*-deficient haematopoietic stem cells (Fig. 3k).

Gene set enrichment analysis (GSEA) revealed significant enrichment of cell-cycle genes and genes with E2F1 motifs in haematopoietic stem cells from mice treated with E2 for 1 week (Extended Data Fig. 8a, b). ERα signalling may therefore promote haematopoietic stem-cell division by activating E2Fs.

Oestrogen levels increase in females during ovulation and pregnancy[28]. Relative to normal female mice, pregnant mice exhibited significantly increased cellularity, erythropoiesis and myelopoiesis in the spleen (Fig. 4a, b) as well as more haematopoietic stem cells in the bone marrow and spleen (Fig. 4c). A 24-h pulse of BrdU to pregnant mothers on day 13.5 of gestation revealed significant increases in proliferation among haematopoietic stem cells, whole bone marrow cells and splenocytes in pregnant as compared to normal female mice (Fig. 4d). As expected[28], serum E2 levels increased significantly in pregnant as compared to control mice (Fig. 4e).

Deletion of *Esr1* from haematopoietic cells in *Mx1-cre; Esr1*$^{fl/fl}$ mice significantly reduced cellularity (Fig. 4a), erythropoiesis (Fig. 4b) and haematopoietic stem-cell numbers (Fig. 4c) in the spleens of pregnant mice relative to pregnant *Esr1*$^{fl/fl}$ controls. Deletion of *Esr1* from haematopoietic cells also significantly reduced BrdU incorporation into haematopoietic stem cells in the bone marrow and spleen of pregnant mice (Fig. 4d). *Esr1* deletion from haematopoietic cells in pregnant mice did not block the increase in haematopoietic stem-cell frequency in the bone marrow but nearly eliminated the increase in haematopoietic stem-cell frequency in the spleen (Fig. 4c). This indicates that oestrogen is not the only factor that increases haematopoietic stem-cell activity in pregnant mice but that it is critical for the mobilization of proliferating haematopoietic stem cells to the spleen and for the expansion of splenic erythropoiesis.

The increase in spleen cellularity and erythropoiesis during pregnancy may also occur in humans, which exhibit increased spleen size during pregnancy[29,30]. There may be many unexplored mechanisms by which systemic signals modulate the function of stem cells within individual tissues in response to physiological change.

METHODS SUMMARY

Specific mouse alleles used in this study are referenced in Methods. Mice were housed in AAALAC-accredited, specific-pathogen-free animal care facilities at the University of Michigan (UM), Baylor College of Medicine (BCM), or UT Southwestern Medical Center (UTSW). All procedures were approved by the UM, BCM and UTSW Institutional Animal Care and Use Committees. For hormonal treatment, mice were injected subcutaneously with 100 µl of corn oil containing 2 µg oestradiol (Sigma), 1 mg progesterone (Sigma)[5], or 100 µg of dihydrotestosterone (Steraloids). 50 µg of anastrozole (Sigma) dissolved in PBS was given intraperitoneally. RU486, PPT and DPN (all from Sigma) dissolved in corn oil were administered subcutaneously at 5 mg kg^{-1}. Poly(I:C) (Amersham) was re-suspended in PBS at 50 µg ml^{-1}, and 0.5 µg g^{-1} of body mass was injected intraperitoneally every other day for 6 days. Female mice were mated with male mice 1 week after the last injection.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 September 2012; accepted 2 December 2013.

1. Williams, T. M. & Carroll, S. B. Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nature Rev. Genet.* **10,** 797–804 (2009).
2. McCarthy, M. M. & Arnold, A. P. Reframing sexual differentiation of the brain. *Nature Neurosci.* **14,** 677–683 (2011).
3. Oatley, J. M. & Brinster, R. L. Regulation of spermatogonial stem cell self-renewal in mammals. *Annu. Rev. Cell Dev. Biol.* **24,** 263–286 (2008).
4. Asselin-Labat, M. L. *et al.* Control of mammary stem cell function by steroid hormone signalling. *Nature* **465,** 798–802 (2010).
5. Joshi, P. A. *et al.* Progesterone induces adult mammary stem cell expansion. *Nature* **465,** 803–807 (2010).
6. Shingo, T. *et al.* Pregnancy-stimulated neurogenesis in the adult female forebrain mediated by prolactin. *Science* **299,** 117–120 (2003).
7. Carreras, E. *et al.* Estradiol acts directly on bone marrow myeloid progenitors to differentially regulate GM-CSF or Flt3 ligand-mediated dendritic cell differentiation. *J. Immunol.* **180,** 727–738 (2008).
8. Thurmond, T. S. *et al.* Role of estrogen receptor alpha in hematopoietic stem cell development and B lymphocyte maturation in the male mouse. *Endocrinology* **141,** 2309–2318 (2000).
9. Medina, K. L. *et al.* Identification of very early lymphoid precursors in bone marrow and their regulation by estrogen. *Nature Immunol.* **2,** 718–724 (2001).
10. Fish, E. N. The X-files in immunity: sex-based differences predispose immune responses. *Nature Rev. Immunol.* **8,** 737–744 (2008).
11. Nakada, D., Levi, B. P. & Morrison, S. J. Integrating physiological regulation with stem cell and tissue homeostasis. *Neuron* **70,** 703–718 (2011).
12. O'Brien, L. E., Soliman, S. S., Li, X. & Bilder, D. Altered modes of stem cell division drive adaptive intestinal growth. *Cell* **147,** 603–614 (2011).
13. Yilmaz, O. H. *et al.* mTORC1 in the Paneth cell niche couples intestinal stem-cell function to calorie intake. *Nature* **486,** 490–495 (2012).
14. McLeod, C. J., Wang, L., Wong, C. & Jones, D. L. Stem cell dynamics in response to nutrient availability. *Curr. Biol.* **20,** 2100–2105 (2010).
15. LaFever, L. & Drummond-Barbosa, D. Direct control of germline stem cell division and cyst growth by neural insulin in *Drosophila*. *Science* **309,** 1071–1073 (2005).
16. Sousa-Nunes, R., Yee, L. L. & Gould, A. P. Fat cells reactivate quiescent neuroblasts via TOR and glial insulin relays in *Drosophila*. *Nature* **471,** 508–512 (2011).
17. Chell, J. M. & Brand, A. H. Nutrition-responsive glia control exit of neural stem cells from quiescence. *Cell* **143,** 1161–1173 (2010).
18. Gourdy, P. *et al.* Relevance of sexual dimorphism to regulatory T cells: estradiol promotes IFN-γ production by invariant natural killer T cells. *Blood* **105,** 2415–2420 (2005).
19. Ghazeeri, G., Abdullah, L. & Abbas, O. Immunological differences in women compared with men: overview and contributing factors. *Am. J. Reprod. Immunol.* **66,** 163–169 (2011).
20. Blobel, G. A. & Orkin, S. H. Estrogen-induced apoptosis by inhibition of the erythroid transcription factor GATA-1. *Mol. Cell. Biol.* **16,** 1687–1694 (1996).
21. Blobel, G. A., Sieff, C. A. & Orkin, S. H. Ligand-dependent repression of the erythroid transcription factor GATA-1 by the estrogen receptor. *Mol. Cell. Biol.* **15,** 3147–3153 (1995).
22. Oguro, H., Ding, L. & Morrison, S. J. SLAM family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell Stem Cell* **13,** 102–116 (2013).
23. Foudi, A. *et al.* Analysis of histone 2B-GFP retention reveals slowly cycling hematopoietic stem cells. *Nature Biotechnol.* **27,** 84–90 (2008).
24. Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135,** 1118–1129 (2008).
25. Yamamoto, R. *et al.* Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. *Cell* **154,** 1112–1126 (2013).
26. Wit, J. M., Hero, M. & Nunez, S. B. Aromatase inhibitors in pediatrics. *Nature Rev. Endocrinol.* **8,** 135–147 (2011).
27. Nilsson, S., Koehler, K. F. & Gustafsson, J. A. Development of subtype-selective oestrogen receptor-based therapeutics. *Nature Rev. Drug Discov.* **10,** 778–792 (2011).
28. Mahendroo, M. S., Cala, K. M., Landrum, D. P. & Russell, D. W. Fetal death in mice lacking 5α-reductase type 1 caused by estrogen excess. *Mol. Endocrinol.* **11,** 917–927 (1997).
29. Sheehan, H. L. & Falkiner, N. M. Splenic aneurysm and splenic enlargement in pregnancy. *BMJ* **2,** 1105 (1948).
30. Maymon, R. *et al.* Normal sonographic values of maternal spleen size throughout pregnancy. *Ultrasound Med. Biol.* **32,** 1827–1831 (2006).

Acknowledgements S.J.M. is a Howard Hughes Medical Institute Investigator and the Mary McDermott Cook Chair in Pediatric Genetics. This work was supported by the Cancer Prevention and Research Institute of Texas (awards to D.N. and S.J.M.) and the National Heart Lung and Blood Institute (HL097760 to S.J.M.). B.P.L. was supported by an Irvington Institute-Cancer Research Institute/Edmond J. Safra Memorial Fellowship. Flow-cytometry was partially supported by the National Institutes of Health (NCRR grant S10RR024574, NIAID AI036211 and NCI P30CA125123) for the BCM Cytometry and Cell Sorting Core. We also thank J. Richards, S. Mani and former members of the Nakada laboratory for discussions. This work was initiated in the Life Sciences Institute at the University of Michigan then completed at Baylor College of Medicine and Children's Research Institute at UT Southwestern. We thank the University of Virginia Center for Research in Reproduction for measuring serum hormone levels. This work is dedicated to Nicole Ryan who passed away during the study.

Author Contributions D.N., H.O. and B.P.L. designed and performed most experiments. G.R.W., A.K., N.R., Y.S. and M.T. performed some experiments with D.N. D.N. and S.J.M. analysed results and wrote the manuscript.

Author Information Microarray data have been deposited to the Gene Expression Omnibus under accession number GSE52711. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.N. (nakada@bcm.edu) or S.J.M. (Sean.Morrison@utsouthwestern.edu).

METHODS

Mice. The mouse alleles used in this study were *Rosa26-rtTA/tetO-H2B-GFP* (ref. 23), germline *Esr1*-deficient[31], *Mx1-cre* (ref. 32), *Vav1-cre* (ref. 33), and *Esr1*^{fl} (ref. 34). Most studies of haematopoietic stem-cell frequency and cycling used young adult C57BL/Ka-Thy-1.1 (CD45.2) mice (8–12 weeks of age). C57BL/Ka-Thy-1.2 (CD45.1) mice were used in transplantation experiments. Mice were housed in AAALAC-accredited, specific-pathogen-free animal care facilities at the University of Michigan (UM), at Baylor College of Medicine (BCM), or UT Southwestern Medical Center (UTSW). All procedures were approved by the UM, BCM and UTSW Institutional Animal Care and Use Committees.

Hormone and poly(I:C) treatments. Mice were injected subcutaneously with 100 µl of corn oil containing 2 µg oestradiol (Sigma) and/or 1 mg progesterone (Sigma)[5]. 100 µg of dihydrotestosterone (Steraloids) in corn oil was administered subcutaneously[35]. 50 µg of anastrozole (Sigma) dissolved in PBS was given intraperitoneally. RU486, PPT and DPN (all from Sigma) dissolved in corn oil were administered subcutaneously at 5 mg kg^{-1}. Poly(I:C) (Amersham) was re-suspended in PBS at 50 µg ml^{-1}, and mice were injected intraperitoneally with 0.5 µg g^{-1} of body mass every other day for 6 days. Note that the biological effect of poly(I:C) varies with polymer length and manufacturer such that doses must be optimized with each batch to obtain complete recombination without inducing haematopoietic stem-cell cycling. Females were mated with male mice 1 week after the last injection.

Statistical methods. Multiple independent experiments were performed to verify the reproducibility of all experimental findings. Group data always represent mean ± standard deviation. Unless otherwise indicated, two-tailed Student's *t*-tests were used to assess statistical significance. No randomization or blinding was used in any experiments. Experimental mice were not excluded from analysis in any experiments. Sample sizes were selected on the basis of previous experience with the degree of variance in each assay.

Cell-cycle analysis. BrdU incorporation *in vivo* was measured by flow cytometry using the APC BrdU Flow Kit (BD Biosciences). Mice were given an intraperitoneal injection of 1 mg of BrdU (Sigma) per 6 g of body mass in PBS and maintained on 1 mg ml^{-1} BrdU in the drinking water for up to 10 days.

Flow cytometry and haematopoietic stem-cell isolation. Bone marrow cells were either flushed from the long bones (tibias and femurs) or isolated by crushing the long bones (tibias and femurs), pelvic bones and vertebrae with mortar and pestle in Hank's buffered salt solution (HBSS) without calcium and magnesium, supplemented with 2% heat-inactivated bovine serum (Gibco). Cells were triturated and filtered through nylon screen (100 µm, Sefar America) or a 40 µm cell strainer (Fisher Scientific) to obtain a single-cell suspension. For isolation of CD150$^+$CD48$^-$Lin$^-$Sca-1$^+$c-kit$^+$ haematopoietic stem cells, bone marrow cells were incubated with PE-Cy5-conjugated anti-CD150 (TC15-12F12.2; BioLegend), PE-conjugated anti-CD48 (HM48-1; BioLegend), APC-conjugated anti-Sca-1 (Ly6A/E; E13-6.7), and biotin-conjugated anti-c-kit (2B8) antibody, in addition to antibodies against the following FITC-conjugated lineage markers: CD41 (MWReg30; BD Biosciences), Ter119, B220 (6B2), Gr1 (8C5), CD2 (RM2-5), CD3 (KT31.1) and CD8 (53-6.7). For isolation of CD34$^-$CD16/32$^-$Lin$^-$Sca-1$^-$c-kit$^+$ MEPs, CD34$^+$CD16/32$^-$Lin$^-$Sca-1$^-$c-kit$^+$ CMPs, and CD34$^+$CD16/32$^+$Lin$^-$Sca-1$^-$c-kit$^+$ GMPs, bone marrow cells were incubated with FITC-conjugated anti-CD34 (RAM34; eBiosciences), PE-Cy7 conjugated anti-CD16/32 (93; Biolegend), PE-Cy5-conjugated anti-Sca-1 (Ly6A/E; E13-6.7), and biotin-conjugated anti-c-kit (2B8) antibody, in addition to antibodies against the following PE-conjugated lineage markers: Ter119, B220 (6B2), Gr1 (8C5), Mac1 (M1/70), CD2 (RM2-5), CD3 (KT31.1) and CD8 (53-6.7). For isolation of Flt3$^+$IL-7R$^+$Lin$^-$Sca-1^{low}c-kit^{low} CLPs, bone marrow cells were incubated with PE-Cy5-conjugated anti-Flt3 (A2F10; eBiosciences), PE-conjugated anti-IL-7R (A7R34; eBiosciences), PE-Cy7-conjugated anti-Sca-1 (Ly6A/E; E13-6.7), and biotin-conjugated anti-c-kit (2B8) antibody, in addition to antibodies against the following FITC-conjugated lineage markers: Ter119, B220 (6B2), Gr1 (8C5), Mac1 (M1/70), CD2 (RM2-5), CD3 (KT31.1) and CD8 (53-6.7). Unless otherwise noted, antibodies were obtained from BioLegend, BD Biosciences, or eBioscience. Biotin-conjugated antibodies were visualized using streptavidin-conjugated APC-Cy7. Haematopoietic stem cells were sometimes pre-enriched by selecting c-kit$^+$ cells using paramagnetic microbeads and autoMACS (Miltenyi Biotec). Nonviable cells were excluded from sorts and analyses using the viability dye 4′,6-diamidino-2-phenylindole (DAPI) (1 µg ml^{-1}). To analyse BrdU incorporation into haematopoietic stem cells, bone marrow cells were incubated with PE-conjugated anti-CD150 (TC15-12F12.2), FITC-conjugated anti-CD48 (HM48-1), PerCP-Cy5.5-conjugated anti-Sca-1 (Ly6A/E; E13-6.7), and biotin-conjugated anti-c-kit (2B8) antibody, in addition to antibodies against the following FITC-conjugated lineage markers: CD41 (MWReg30), Ter119, B220 (6B2), Gr1 (8C5), CD2 (RM2-5), CD3 (KT31.1) and CD8 (53-6.7). Biotin-conjugated c-kit was visualized using streptavidin-conjugated AlexaFluor 700 or PE-Cy7 (Invitrogen). To distinguish donor haematopoietic stem cells from recipient haematopoietic stem cells, AlexaFluor 700 conjugated anti-CD45.2 (clone 104) antibody was used. To analyse haematopoietic lineage composition, bone marrow cells or splenocytes were incubated with FITC-conjugated anti-B220, PE-conjugated anti-Ter119, APC-conjugated anti-CD3, APC-eFluor780-conjugated anti-Mac1 (M1/70), and PE-Cy7-conjugated anti-Gr1 antibodies. Annexin-V staining was performed using Annexin-V APC (BD Biosciences). Flow cytometry was performed with FACSAria II, FACSCanto II, LSR II, or LSRFortessa flow-cytometers (BD Biosciences).

Ki67 staining. Ki67 staining was performed as before[36]. In brief, haematopoietic stem cells were sorted into 70% ethanol and kept at −20 °C for at least 24 h. Ki67 staining was performed using a FITC Ki-67 kit (BD Biosciences), followed by staining with 50 µg ml^{-1} propidium iodide (Molecular Probes) and analysed by flow cytometry.

Castration. An incision was made in the scrotum and the testis and attached testicular fat pads were pulled out of the incision. Spermatochords were individually ligated with absorbable sutures (4-0 chromic gut), then excised, and then 1–3 non-absorbable sutures (3-0 Tevdek II) were used to close the skin.

Ovariectomy. The skin around the dorsal midline caudal to the posterior borders of the ribs was shaved and an incision was made to expose the ovaries on each side. The ovaries were isolated, ligated with absorbable sutures (4-0 chromic gut) and excised, and then 3–4 non-absorbable sutures (3-0 Tevdek II) were used to close the skin. Sham-treated mice underwent similar surgeries except that the gonads were left intact. All animals were allowed to recover for 2 weeks before BrdU was administered.

Cell-cycle analysis of haematopoietic stem cells from competitively reconstituted mice. Adult recipient mice (CD45.1) were irradiated with an X-ray source delivering approximately 300 rad min^{-1} in two equal doses of 540 rad, delivered at least 2 h apart. 10^6 whole bone marrow cells from CD45.2 *Vav1-cre; Esr1*$^{fl/fl}$ mice were transplanted along with 10^6 CD45.1 whole bone marrow cells from wild-type mice into the retro-orbital venous sinus of anaesthetized recipients. Fifteen weeks after transplantation, recipient mice were either treated with oil or 2 µg oestradiol for 7 days. Mice were administered BrdU continuously for the last 3 days of oestradiol treatment. BrdU incorporation into haematopoietic stem cells was analysed by flow cytometry as described above.

Measurement of serum hormone concentration. Whole blood samples were collected and allowed to clot at room temperature for 90 min before being centrifuged at 2,000*g* for 15 min at room temperature. Serum samples were analysed for oestradiol and progesterone levels at the University of Virginia Center for Research in Reproduction.

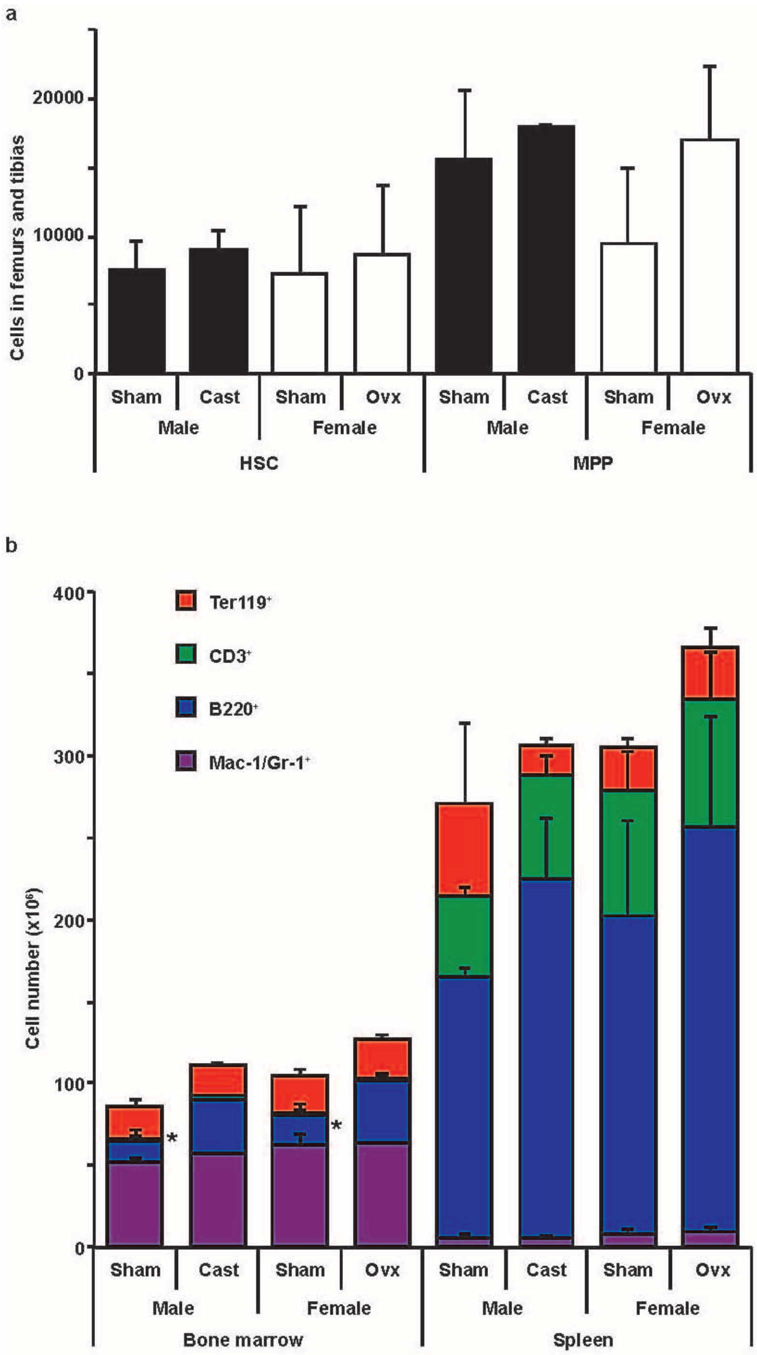
Quantitative real-time (reverse transcription) PCR. Haematopoietic stem cells and other haematopoietic cells were sorted into Trizol (Life Technologies) and RNA was isolated according to the manufacturer's instructions. cDNA was made with random primers and SuperScript III reverse transcriptase (Life Technologies). Quantitative PCR was performed using a LightCycler 480 (Roche Applied Science) or ViiA7 Real-Time PCR System (Life Technologies). Each sample was normalized to β-actin. Primers to quantify cDNA levels were *Esr1* forward, 5′-CCTTCTAGA CCCTTCAGTGAAGCC-3′, *Esr1* reverse, 5′-CGAGACCAATCATCAGAATC TCC-3′; *Esr2* forward, 5′- CCAGCCCTGTTACTAGTCCAAGC-3′, *Esr2* reverse, 5′-GGTACACTGATTCGTGGCTGG-3′, *Pgr* forward, 5′-CCAGCTCACAGCG CTTCTACC-3′, *Pgr* reverse, 5′- GAAAGAGGAGCGGCTTCACC-3′, *Ar* forward, 5′-GGTGTGTGCCGGACATGACAAC-3′, *Ar* reverse, 5′-GGTCATCCA CATGCAAGTTGCGG-3′, β-actin forward, 5′- CGTCGACAACGGCTCCGG CATG-3′ and β-actin reverse, 5′- GGGCCTCGTCACCCACATAGGAG-3′.

Microarray analysis. Groups of three male and three female mice were treated with either E2 (2 µg per day) or vehicle (oil) for one week. Haematopoietic stem cells were sorted into Trizol and RNA purified according to the manufacturer's instructions. DNase-I-treated RNA samples were amplified and biotinylated using the Nugen Ovation Pico WTA V2 system and the Encore Biotin Module (NuGEN Technologies). Biotinylated samples were hybridized to Affymetrix Mouse Gene 1.0 ST Arrays (Affymetrix) by the Baylor College of Medicine Genomic and RNA profiling core. dChip software[37] was used to calculate normalized expression values for each array. Gene set enrichment analysis was performed as described previously[38].

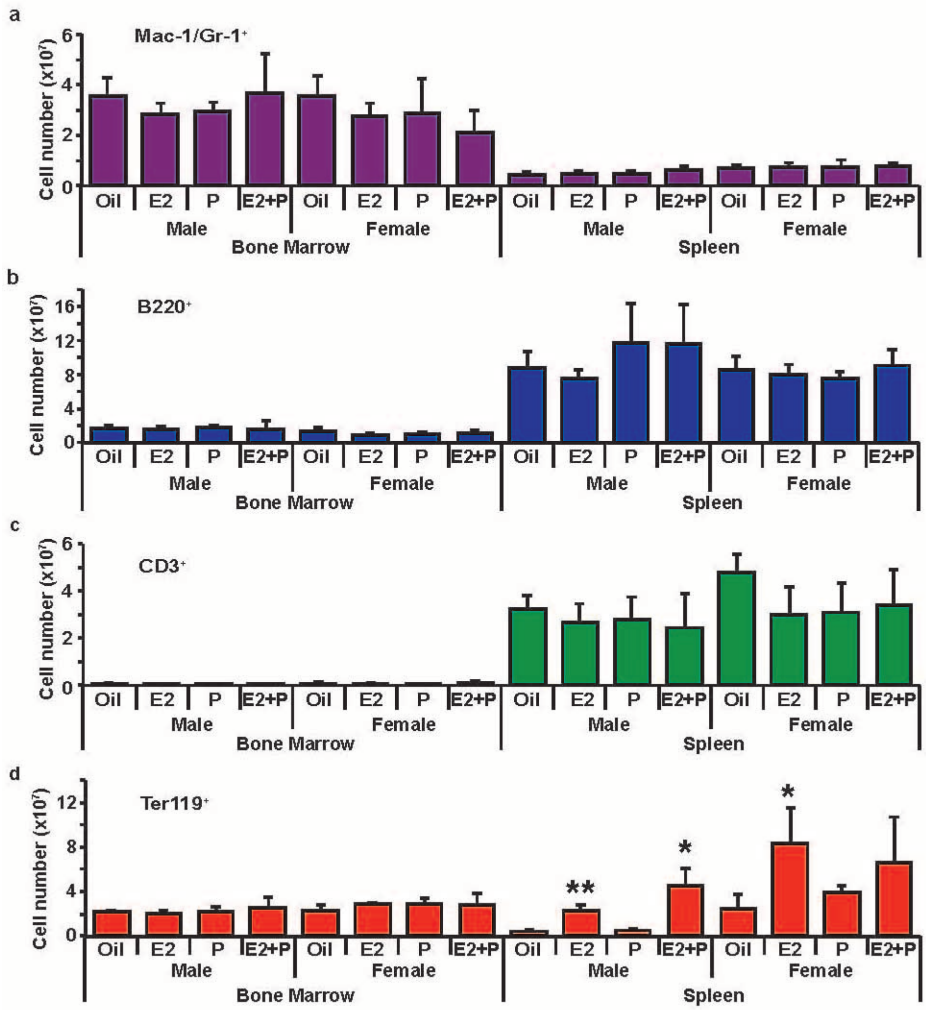
Haematopoietic stem-cell culture and *in vitro* BrdU incorporation. Haematopoietic stem cells were sorted directly into serum-free, phenol-red-free medium (X-Vivo 15, Lonza) supplemented with 50 ng ml^{-1} of SCF and 50 ng ml^{-1} TPO (both from Peprotech), with or without 100 nM oestradiol, and cultured for 3 days. BrdU (10 µM final concentration) was added for an hour before cells were cytospun to a slide. Slides were fixed with cold methanol for 5 min at −20 °C, then washed with PBS containing 0.01% NP-40 and treated with 2N HCl for 15 min. Slides were blocked in PBS containing 4% goat serum, 4 mg ml^{-1} BSA, and 0.1% NP-40 followed by staining overnight at 4 °C with antibody against BrdU (BU1/75, 1:100, Abcam) as described previously[36].

Immunostaining. Sorted haematopoietic stem cells were fixed with methanol and stained overnight at 4 °C with antibodies against ERα (MC-20, 1:500, Santa Cruz Biotechnology) diluted in PBS containing 4% goat serum, $4\,\mathrm{mg\,ml^{-1}}$ BSA and 0.1% NP-40. Primary antibody staining was developed with secondary antibodies conjugated to Alexa fluor 488 together with DAPI ($2\,\mu\mathrm{g\,ml^{-1}}$). Slides were analysed on a Leica DMI6000 fluorescence microscope.

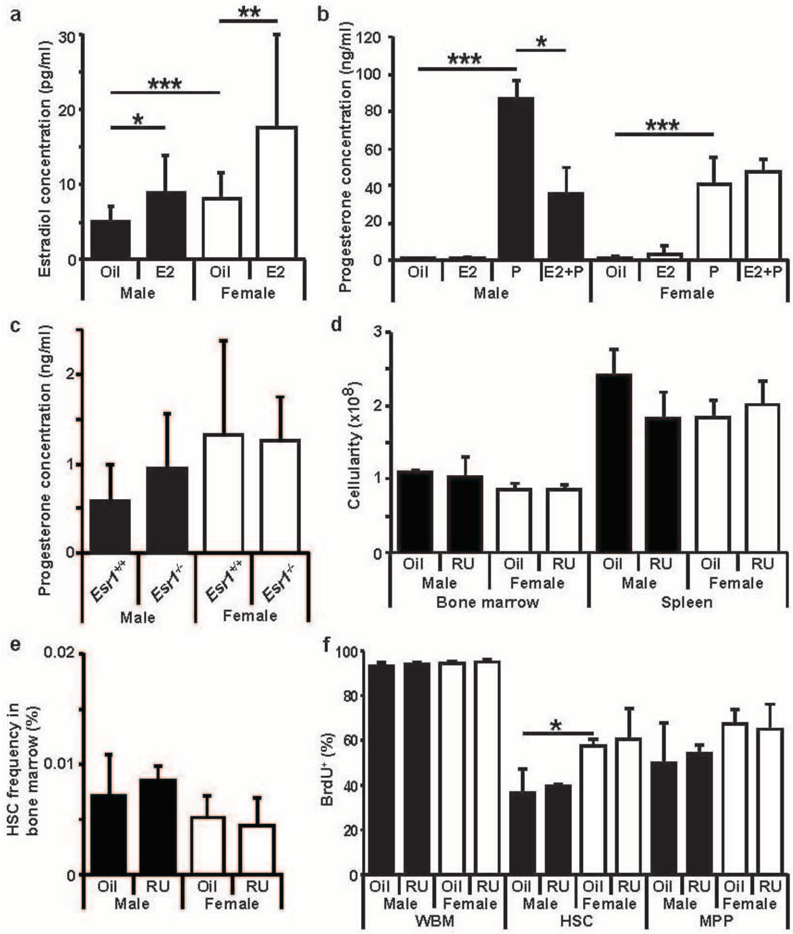
31. Lubahn, D. B. *et al.* Alteration of reproductive function but not prenatal sexual development after insertional disruption of the mouse estrogen receptor gene. *Proc. Natl Acad. Sci. USA* **90,** 11162–11166 (1993).
32. Kühn, R., Schwenk, F., Aguet, M. & Rajewsky, K. Inducible gene targeting in mice. *Science* **269,** 1427–1429 (1995).
33. de Boer, J. *et al.* Transgenic mice with hematopoietic and lymphoid specific expression of Cre. *Eur. J. Immunol.* **33,** 314–325 (2003).
34. Feng, Y., Manka, D., Wagner, K. U. & Khan, S. A. Estrogen receptor-α expression in the mammary epithelium is required for ductal and alveolar morphogenesis in mice. *Proc. Natl Acad. Sci. USA* **104,** 14718–14723 (2007).
35. Azzi, L., El-Alfy, M., Martel, C. & Labrie, F. Gender differences in mouse skin morphology and specific effects of sex steroids and dehydroepiandrosterone. *J. Invest. Dermatol.* **124,** 22–27 (2005).
36. Nakada, D., Saunders, T. L. & Morrison, S. J. Lkb1 regulates cell cycle and energy metabolism in haematopoietic stem cells. *Nature* **468,** 653–658 (2010).
37. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA* **98,** 31–36 (2001).
38. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).



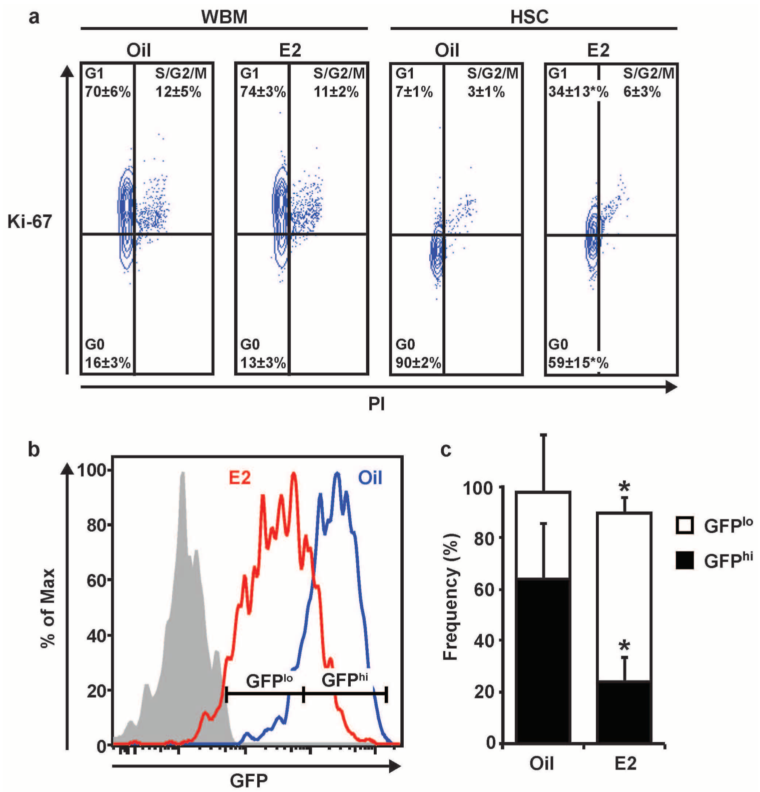
Extended Data Figure 1 | Castration or ovariectomy modestly increased the numbers of B cells in the bone marrow without affecting the numbers of haematopoietic stem cells or MPPs. **a**, Castration (cast) or ovariectomy (ovx) did not significantly affect the numbers of haematopoietic stem cells or MPPs in the bone marrow (femurs and tibias). **b**, Castration or ovariectomy significantly increased the numbers of $B220^+$ B cells in the bone marrow but did not affect the numbers of $Mac1^+/Gr1^+$ myeloid cells, $CD3^+$ T cells, or $Ter119^+$ erythroid cells in the bone marrow or spleen. 3 sham and 4 gonadectomized mice used in 3 independent experiments. All data represent mean ± standard deviation; $^*P < 0.05$ by Student's t-test.



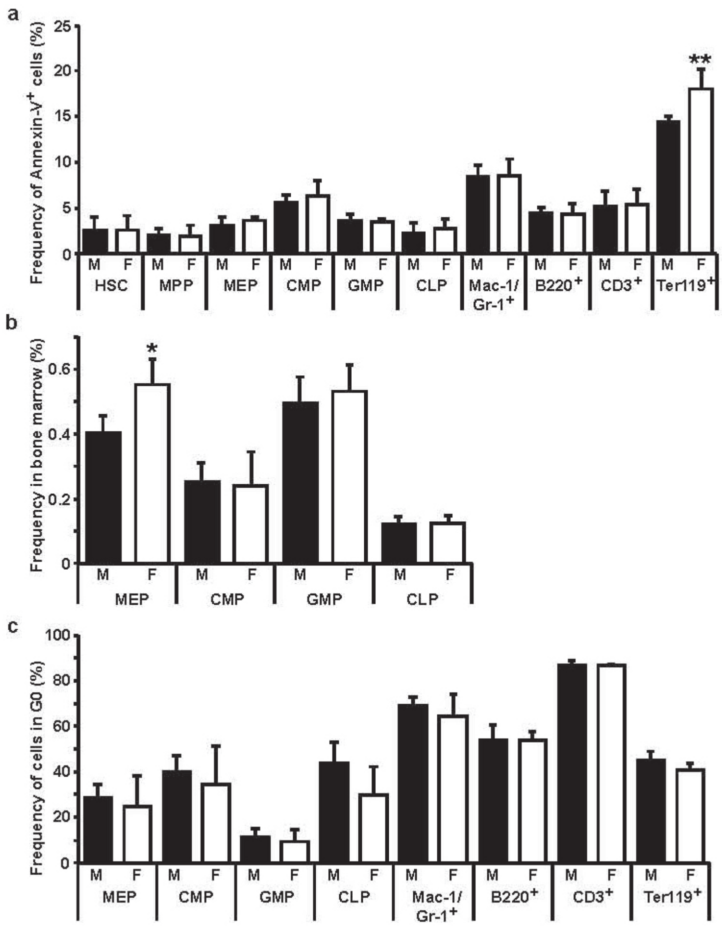
Extended Data Figure 2 | Administration of oestradiol (E2) to mice induced erythropoiesis in the spleen. a–c, Treatment of male and female mice for 1 week with E2, with or without P, did not affect the numbers of Mac1$^+$/Gr1$^+$ myeloid cells, B220$^+$ B cells, or CD3$^+$ T cells in the bone marrow or spleen of either sex. **d**, E2 and E2+P treatment did significantly increase the number of Ter119$^+$ erythroid cells in the spleen of male mice, and E2 treatment significantly increased the number of Ter119$^+$ erythroid cells in the spleen of female mice. $n = 3$ mice per treatment in 3 independent experiments. All data represent mean ± standard deviation; $*P < 0.05$; $**P < 0.005$; $***P < 0.0005$ by Student's t-test comparing each treatment to vehicle.



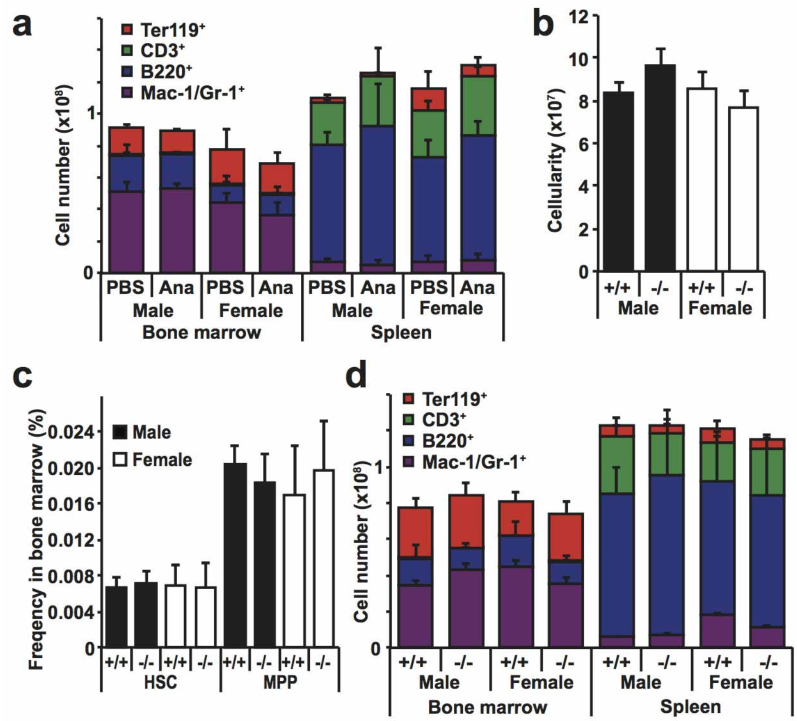
Extended Data Figure 3 | Administration of exogenous oestrogen and progesterone significantly increased serum oestrogen and progesterone levels in mice but progesterone did not affect haematopoietic stem-cell division *in vivo*. **a**, Oestradiol treatment significantly increased serum oestradiol levels in male and female mice but the increased levels remained within the physiological range, similar to levels observed during pregnancy (see Fig. 4e) (male oil, 22; male E2, 20; female oil, 33; female E2, 14 mice used in 8 independent experiments). **b**, Progesterone treatment significantly increased serum progesterone levels in male and female mice ($n = 3$ mice per treatment in 3 independent experiments). Note that this did not affect bone marrow or spleen cellularity, haematopoietic stem-cell frequency, or haematopoietic stem-cell division (Fig. 2b–d). **c**, *Esr1*-deficient mice had normal levels of serum progesterone ($n = 3$ mice per group in 3 independent experiments). **d–f**, Administration of a progesterone receptor antagonist, RU486 (RU), did not affect bone marrow or spleen cellularity (**d**), haematopoietic stem-cell frequency in the bone marrow (**e**), or the division of haematopoietic stem cells, MPPs, or WBM cells (**f**). All data represent mean ± standard deviation from 3 independent experiments, except as indicated above; *$P < 0.05$; **$P < 0.005$; ***$P < 0.0005$ by Student's *t*-test comparing each treatment to vehicle (oil).



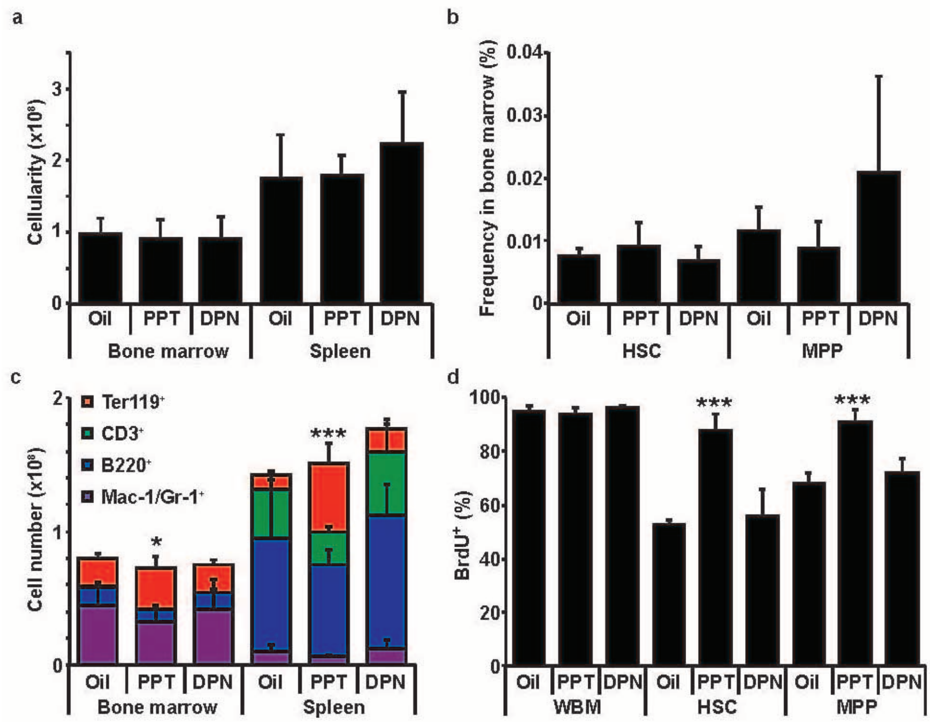
Extended Data Figure 4 | Oestrogen treatment increased the frequency of Ki67$^+$ cycling haematopoietic stem cells. **a**, Administration of oestrogen (E2) significantly increased the frequency of haematopoietic stem cells in G1 phase of the cell cycle, and reduced the frequency of haematopoietic stem cells in G0 phase of the cell cycle based on Ki67/propidium iodide staining ($n = 3$ mice per treatment in 3 independent experiments). **b**, **c**, *Col1A1-H2B-GFP; Rosa26-M2-rtTA* mice pulsed with doxycycline for 6 weeks to induce H2B–GFP expression were treated with oil (blue histogram) or E2 (red histogram) for 2 weeks without doxycycline. E2 treatment significantly increased the rate of haematopoietic stem-cell division as indicated by the reduced frequency of GFP^{high} quiescent haematopoietic stem cells and the increased frequency of GFP^{low} moderately cycling haematopoietic stem cells (5 oil-treated and 4 E2-treated mice used in 3 independent experiments). All data represent mean ± standard deviation; *$P < 0.05$ by Student's t-test comparing each treatment to vehicle (oil).



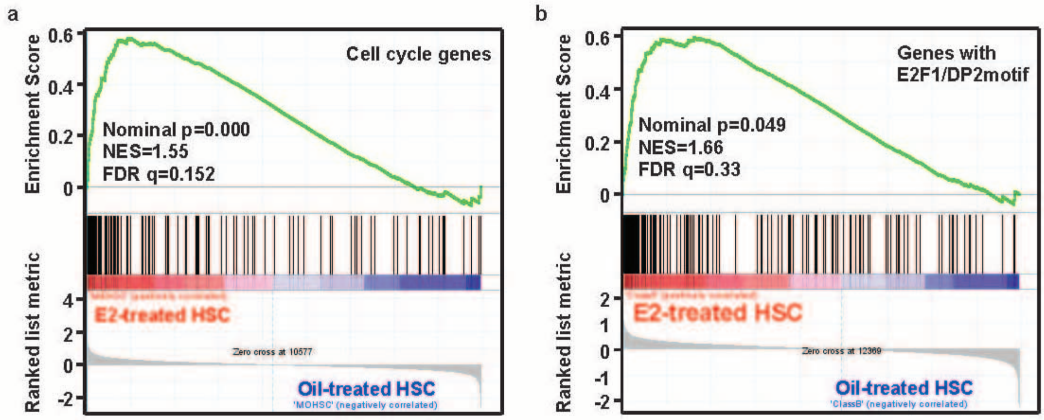
Extended Data Figure 5 | Female mice have increased frequencies of megakaryocyte–erythroid progenitors (MEPs) and apoptotic Ter119$^+$ cells relative to male mice. **a**, Annexin-V staining of the indicated cell populations in the bone marrow of male and female mice revealed a significantly increased frequency of apoptotic Annexin-V$^+$ cells among Ter119$^+$ erythroid progenitors in female mice. **b**, Female mice had a significantly increased frequency of CD34$^-$CD16/32$^-$Lin$^-$Sca-1$^-$c-kit$^+$ MEPs, but no significant differences in the frequencies of CD34$^+$CD16/32$^-$Lin$^-$Sca-1$^-$c-kit$^+$ CMPs, CD34$^+$CD16/32$^+$Lin$^-$Sca-1$^-$c-kit$^+$ GMPs, or Flt3$^+$IL-7R$^+$Lin$^-$Sca-1^{low}c-kit^{low} CLPs. **c**, None of the restricted progenitors or differentiated cells displayed differences in cell-cycle status between male and female mice (**a**–**c**, $n = 5$ mice per group in three independent experiments). Data represent mean ± standard deviation; *$P < 0.05$; **$P < 0.005$; by Student's *t*-test comparing each treatment between sexes.



Extended Data Figure 6 | Inhibiting oestrogen signalling by anastrozole treatment or *Esr1* deficiency did not affect the numbers of haematopoietic cells in the bone marrow or spleen. **a**, Administration of anastrozole (Ana) to mice for 2 weeks did not significantly affect the number of Ter119$^+$ erythroid cells, CD3$^+$ T cells, B220$^+$ B cells, or Mac1$^+$/Gr1$^+$ myeloid cells in the bone marrow or spleen (4 PBS-treated and 6 anastrozole-treated mice were used in 4 independent experiments). **b**, **c**, *Esr1* deficiency did not significantly affect bone marrow cellularity (**b**) or the frequencies of haematopoietic stem cells or MPPs (**c**) in either sex. **d**, *Esr1* deficiency did not significantly affect the numbers of Ter119$^+$ erythroid cells, CD3$^+$ T cells, B220$^+$ B cells, or Mac1$^+$/Gr1$^+$ myeloid cells in the bone marrow or spleen of normal mice. $-/-$ indicates *Esr1*-deficient mice and $+/+$ indicates wild-type littermate control mice (**b**–**d**, $n = 3$ mice per group in 3 independent experiments). All data represent mean $\pm$ standard deviation.



Extended Data Figure 7 | Pharmacological ERα activation, but not ERβ activation, is sufficient to promote haematopoietic stem-cell division. Male mice ($n = 5$ mice per treatment in 3 independent experiments) were treated with oil, ERα agonist PPT, or ERβ agonist DPN for 14 days then pulsed with BrdU for 10 days, beginning on the fourth day of hormone treatment. **a**, **b**, PPT or DPN treatment did not significantly affect the cellularity of bone marrow or spleen (**a**), or the frequencies of haematopoietic stem cells or MPPs in bone marrow (**b**). **c**, PPT treatment, but not DPN treatment, significantly increased erythropoiesis in bone marrow and spleen. **d**, PPT significantly increased the division rates of haematopoietic stem cells and MPPs, but DPN failed to do so, suggesting that ERα activation, but not ERβ activation, promotes haematopoietic stem cell division. Data represent mean ± standard deviation; ***$P < 0.0005$ by Student's t-test comparing each treatment to vehicle (oil).



Extended Data Figure 8 | E2 treatment changes haematopoietic stem-cell gene expression profile in a manner consistent with increased cell division. **a**, **b**, Gene set enrichment analysis revealed that haematopoietic stem cells from mice treated with E2 exhibited significant enrichment in the expression of genes involved in cell cycling (**a**). We also observed a significant enrichment in the expression of genes with E2F1 motifs in haematopoietic stem cells from E2-treated mice, consistent with the role of E2Fs in cell-cycle control (**b**). $n = 3$ mice per treatment of each sex were used to isolate independent aliquots of RNA from haematopoietic stem cells for gene expression profiling.

doi:10.1038/nature12820

Diet rapidly and reproducibly alters the human gut microbiome

Lawrence A. David[1,2]†, Corinne F. Maurice[1], Rachel N. Carmody[1], David B. Gootenberg[1], Julie E. Button[1], Benjamin E. Wolfe[1], Alisha V. Ling[3], A. Sloan Devlin[4], Yug Varma[4], Michael A. Fischbach[4], Sudha B. Biddinger[3], Rachel J. Dutton[1] & Peter J. Turnbaugh[1]

Long-term dietary intake influences the structure and activity of the trillions of microorganisms residing in the human gut[1–5], but it remains unclear how rapidly and reproducibly the human gut microbiome responds to short-term macronutrient change. Here we show that the short-term consumption of diets composed entirely of animal or plant products alters microbial community structure and overwhelms inter-individual differences in microbial gene expression. The animal-based diet increased the abundance of bile-tolerant microorganisms (*Alistipes*, *Bilophila* and *Bacteroides*) and decreased the levels of Firmicutes that metabolize dietary plant polysaccharides (*Roseburia*, *Eubacterium rectale* and *Ruminococcus bromii*). Microbial activity mirrored differences between herbivorous and carnivorous mammals[2], reflecting trade-offs between carbohydrate and protein fermentation. Foodborne microbes from both diets transiently colonized the gut, including bacteria, fungi and even viruses. Finally, increases in the abundance and activity of *Bilophila wadsworthia* on the animal-based diet support a link between dietary fat, bile acids and the outgrowth of microorganisms capable of triggering inflammatory bowel disease[6]. In concert, these results demonstrate that the gut microbiome can rapidly respond to altered diet, potentially facilitating the diversity of human dietary lifestyles.

There is growing concern that recent lifestyle innovations, most notably the high-fat/high-sugar 'Western' diet, have altered the genetic composition and metabolic activity of our resident microorganisms (the human gut microbiome)[7]. Such diet-induced changes to gut-associated microbial communities are now suspected of contributing to growing epidemics of chronic illness in the developed world, including obesity[4,8] and inflammatory bowel disease[6]. Yet, it remains unclear how quickly and reproducibly gut bacteria respond to dietary change. Work in inbred mice shows that shifting dietary macronutrients can broadly and consistently alter the gut microbiome within a single day[7,9]. By contrast, dietary interventions in human cohorts have only measured community changes on timescales of weeks[10] to months[4], failed to find significant diet-specific effects[1], or else have demonstrated responses among a limited number of bacterial taxa[3,5].

We examined whether dietary interventions in humans can alter gut microbial communities in a rapid, diet-specific manner. We prepared two diets that varied according to their primary food source: a 'plant-based diet', which was rich in grains, legumes, fruits and vegetables; and an 'animal-based diet', which was composed of meats, eggs and cheeses (Supplementary Table 1). We picked these sources to span the global diversity of modern human diets, which includes exclusively plant-based and nearly exclusively animal-based regimes[11] (the latter being the case among some high-latitude and pastoralist cultures). Each diet was consumed *ad libitum* for five consecutive days by six male and four female American volunteers between the ages of 21 and 33, whose body mass indices ranged from 19 to 32 $kg\,m^{-2}$ (Supplementary Table 2). Study volunteers were observed for 4 days before each diet arm to measure normal eating habits (the baseline period) and for 6 days after each diet arm to assess microbial recovery (the washout period; Extended Data Fig. 1). Subjects' baseline nutritional intake correlated well with their estimated long-term diet (Supplementary Table 3). Our study cohort included a lifetime vegetarian (see Extended Data Fig. 2, Supplementary Discussion and Supplementary Table 4 for a detailed analysis of his diet and gut microbiota).

Each diet arm significantly shifted subjects' macronutrient intake (Fig. 1a–c). On the animal-based diet, dietary fat increased from $32.5 \pm 2.2\%$ to $69.5 \pm 0.4\%$ kcal and dietary protein increased from $16.2 \pm 1.3\%$ to $30.1 \pm 0.5\%$ kcal ($P < 0.01$ for both comparisons, Wilcoxon signed-rank test; Supplementary Table 5). Fibre intake was nearly zero, in contrast to baseline levels of 9.3 ± 2.1 g per 1,000 kcal. On the plant-based diet, fibre intake rose to 25.6 ± 1.1 g per 1,000 kcal, whereas both fat and protein intake declined to $22.1 \pm 1.7\%$ and $10.0 \pm 0.3\%$ kcal, respectively ($P < 0.05$ for all comparisons). Subjects' weights on the plant-based diet remained stable, but decreased significantly by day 3 of the animal-based diet ($q < 0.05$, Bonferroni-corrected Mann–Whitney U test; Extended Data Fig. 3). Differential weight loss between the two diets cannot be explained simply by energy intake, as subjects consumed equal numbers of calories on the plant- and animal-based diets ($1{,}695 \pm 172$ kcal and $1{,}777 \pm 221$ kcal, respectively; $P = 0.44$).

To characterize temporal patterns of microbial community structure, we performed 16S ribosomal RNA gene sequencing on samples collected each day of the study (Supplementary Table 6). We quantified the microbial diversity within each subject at a given time point (α diversity) and the difference between each subject's baseline and diet-associated gut microbiota (β diversity) (Fig. 1d, e). Although no significant differences in α diversity were detected on either diet, we observed a significant increase in β diversity that was unique to the animal-based diet ($q < 0.05$, Bonferroni-corrected Mann–Whitney U test). This change occurred only 1 day after the diet reached the distal gut microbiota (as indicated by the food tracking dye; Extended Data Fig. 3a). Subjects' gut microbiota reverted to their original structure 2 days after the animal-based diet ended (Fig. 1e).

Analysis of the relative abundance of bacterial taxonomic groups supported our finding that the animal-based diet had a greater impact on the gut microbiota than the plant-based diet (Fig. 2). We hierarchically clustered species-level bacterial phylotypes by the similarity of their dynamics across diets and subjects (see Methods and Supplementary Tables 7, 8). Statistical testing identified 22 clusters whose abundance significantly changed while on the animal-based diet, whereas only 3 clusters showed significant abundance changes while on the plant-based diet ($q < 0.05$, Wilcoxon signed-rank test; Supplementary Table 9). Notably, the genus *Prevotella*, one of the leading sources of inter-individual gut microbiota variation[12] and hypothesized to be sensitive to long-term fibre intake[1,13], was reduced in our vegetarian subject during consumption of the animal-based diet (see Supplementary

[1]FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Society of Fellows, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Division of Endocrinology, Children's Hospital Boston, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Department of Bioengineering & Therapeutic Sciences and the California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, California 94158, USA. †Present address: Molecular Genetics & Microbiology and Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina 27708, USA.

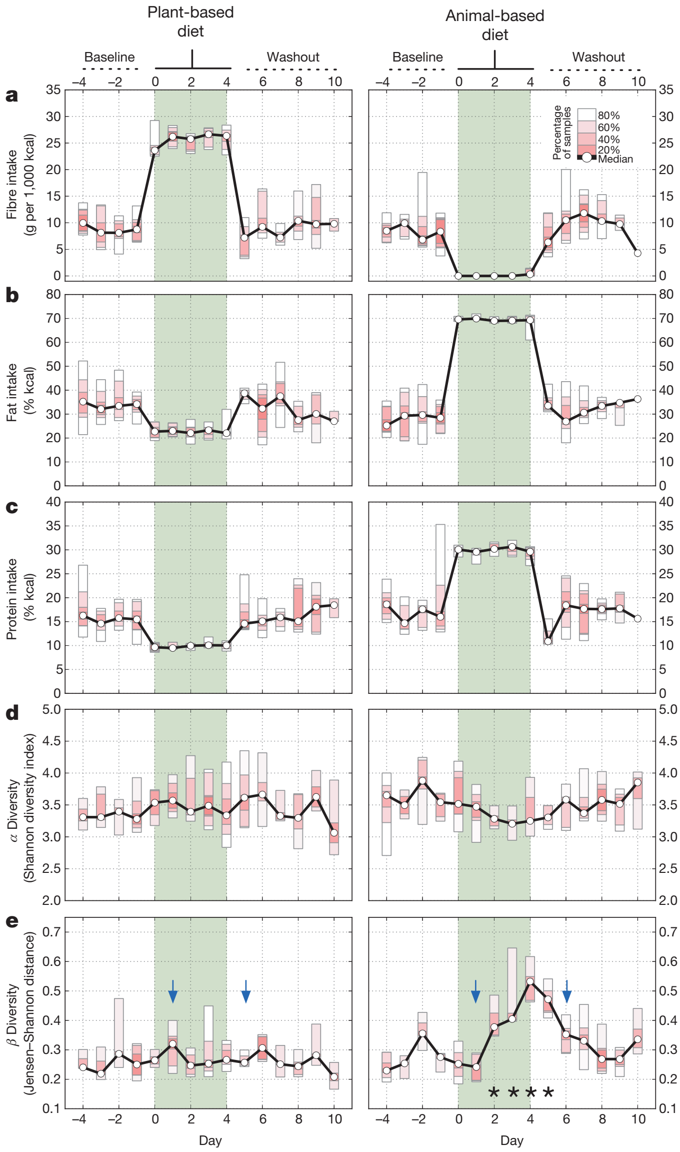


Figure 1 | Short-term diet alters the gut microbiota. a–e, Ten subjects were tracked across each diet arm. **a**, Fibre intake on the plant-based diet rose from a median baseline value of 9.3 ± 2.1 to 25.6 ± 1.1 g per 1,000 kcal ($P = 0.007$; two-sided Wilcoxon signed-rank test), but was negligible on the animal-based diet ($P = 0.005$). **b**, Daily fat intake doubled on the animal-based diet from a baseline of 32.5 ± 2.2% to 69.5 ± 0.4% kcal ($P = 0.005$), but dropped on the plant-based diet to 22.1 ± 1.7% kcal ($P = 0.02$). **c**, Protein intake rose on the animal-based diet to 30.1 ± 0.5% kcal from a baseline level of 16.2 ± 1.3% kcal ($P = 0.005$), and decreased on the plant-based diet to 10.0 ± 0.3% kcal ($P = 0.005$). **d**, Within-sample species diversity (α diversity, Shannon diversity index), did not significantly change during either diet. **e**, The similarity of each individual's gut microbiota to their baseline communities (β diversity, Jensen–Shannon distance) decreased on the animal-based diet (dates with $q < 0.05$ identified with asterisks; Bonferroni-corrected, two-sided Mann–Whitney *U* test). Community differences were apparent 1 day after a tracing dye showed the animal-based diet reached the gut (blue arrows depict appearance of food dyes added to first and last diet day meals; Extended Data Fig. 3a).

Discussion). We also observed a significant positive correlation between subjects' fibre intake over the past year and baseline gut *Prevotella* levels (Extended Data Fig. 4 and Supplementary Table 10).

To identify functional traits linking clusters that thrived on the animal-based diet, we selected the most abundant taxon in the three most-enriched clusters (*Bilophila wadsworthia*, cluster 28; *Alistipes putredinis*, cluster 26; and a *Bacteroides sp.*, cluster 29), and performed a literature search for their lifestyle traits. That search quickly yielded a

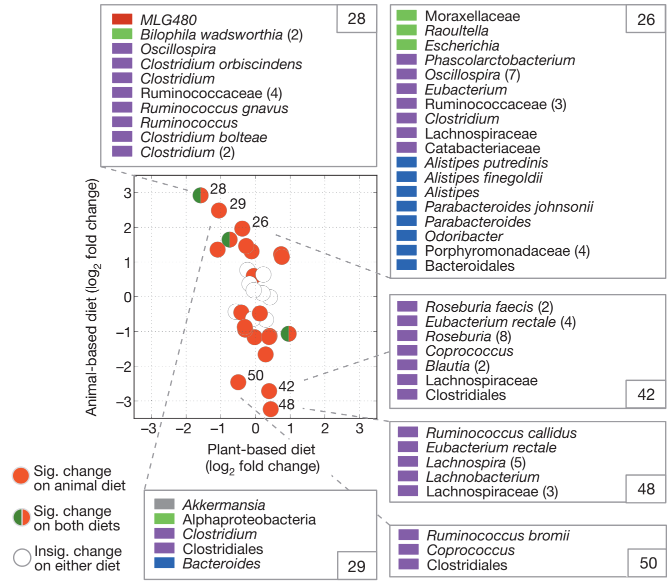


Figure 2 | Bacterial cluster responses to diet arms. Cluster $\log_2$ fold changes on each diet arm were computed relative to baseline samples across all subjects and are drawn as circles. Clusters with significant (Sig.) fold changes on the animal-based diet are coloured in red, and clusters with significant fold changes on both the plant- and animal-based diets are coloured in both red and green. Uncoloured clusters exhibited no significant (Insig.) fold change on either the animal- or plant-based diet ($q < 0.05$, two-sided Wilcoxon signed-rank test). Bacterial membership in the clusters with the three largest positive and negative fold changes on the animal-based diet are also displayed and coloured by phylum: Firmicutes (purple), Bacteroidetes (blue), Proteobacteria (green), Tenericutes (red) and Verrucomicrobia (grey). Multiple operational taxonomic units (OTUs) with the same name are counted in parentheses.

common theme of bile resistance for these taxa, which is consistent with observations that high fat intake causes more bile acids to be secreted[14].

Analysis of faecal short-chain fatty acids (SCFAs) and bacterial clusters suggests that macronutrient shifts on both diets also altered microbial metabolic activity. Relative to the plant-based diet and baseline samples, the animal-based diet resulted in significantly lower levels of the products of carbohydrate fermentation and a higher concentration of the products of amino acid fermentation (Fig. 3a, b and Supplementary Table 11). When we correlated subjects' SCFA concentrations with the same-day abundance of bacterial clusters from Fig. 2, we found significant positive relationships between clusters composed of putrefactive microbes[15,16] (that is, *Alistipes putredinis* and *Bacteroides* spp.) and SCFAs that are the end products of amino acid fermentation (Extended Data Fig. 5). We also observed significant positive correlations between clusters comprised of saccharolytic microbes[3] (for example, *Roseburia*, *E. rectale* and *F. prausnitzii*) and the products of carbohydrate fermentation.

To test whether the observed changes in microbial community structure and metabolic end products were accompanied by more widespread shifts in the gut microbiome, we measured microbial gene expression using RNA sequencing (RNA-seq). A subset of samples was analysed, targeting the baseline periods and the final 2 days of each diet (Extended Data Fig. 1 and Supplementary Table 12). We identified several differentially expressed metabolic modules and pathways during the plant- and animal-based diets (Supplementary Tables 13, 14). The animal-based diet was associated with increased expression of genes for vitamin biosynthesis (Fig. 3c); the degradation of polycyclic aromatic hydrocarbons (Fig. 3d), which are carcinogenic compounds produced during the charring of meat[17]; and the increased expression of β-lactamase genes (Fig. 3e). Metagenomic models constructed from our 16S rRNA data[18] suggest that the observed expression differences are due to a combination of regulatory and taxonomic shifts within the microbiome (Supplementary Tables 15, 16).

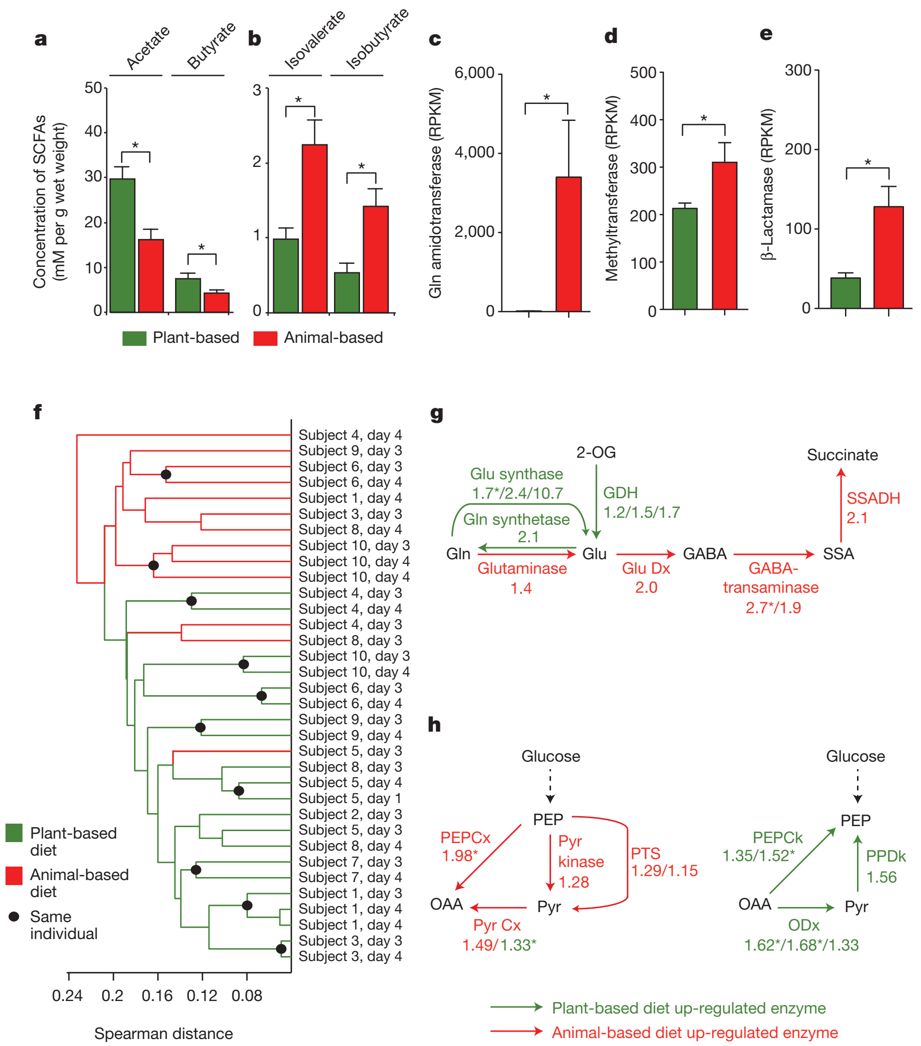


Figure 3 | Diet alters microbial activity and gene expression. **a**, **b**, Faecal concentrations of SCFAs from carbohydrate (**a**) and amino acid (**b**) fermentation (*$P < 0.05$, two-sided Mann–Whitney U test; $n = 9$–11 faecal samples per diet arm; Supplementary Table 11). **c**–**e**, The animal-based diet was associated with significant increases in gene expression (normalized to reads per kilobase per million mapped (RPKM); $n = 13$–21 data sets per diet arm) among glutamine amidotransferases (KEGG orthologous group K08681, vitamin B_6 metabolism) (**c**), methyltransferases (K00599, polycyclic aromatic hydrocarbon degradation) (**d**) and β-lactamases (K01467) (**e**). **f**, Hierarchical clustering of gut microbial gene expression profiles collected on the animal-based (red) and plant-based (green) diets. Expression profile similarity was significantly associated with diet ($P < 0.003$; two-sided Fisher's exact test excluding replicate samples), despite inter-individual variation that preceded the diet (Extended Data Fig. 6a, b). **g**, **h**, Enrichment on animal-based diet (red) and plant-based diet (green) for expression of genes involved in amino acid metabolism (**g**) and central metabolism (**h**). Numbers indicate the mean fold change between the two diets for each KEGG orthologous group assigned to a given enzymatic reaction (Supplementary Table 17). Enrichment patterns on the animal- and plant-based diets agree perfectly with patterns observed in carnivorous and herbivorous mammals, respectively[2] ($P < 0.001$, Binomial test). GDH, glutamate dehydrogenase; Glu Dx, glutamate decarboxylase; ODx, oxaloacetate decarboxylase; PEPCx, phosphoenolpyruvate carboxylase; PEPCk, PEP carboxykinase; PPDk, pyruvate, orthophosphate dikinase; PTS, phosphotransferase system; Pyr Cx, pyruvate carboxylase; SSADH, succinate-semialdehyde dehydrogenase. Note that Pyr Cx is represented by two groups, which showed divergent fold changes. **c**–**h**, *$P < 0.05$, Student's t-test. Values in panels **a**–**e** are mean ± standard error of the mean (s.e.m.).

Next, we hierarchically clustered microbiome samples based on the transcription of Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologous groups[19], which suggested that overall microbial gene expression was strongly linked to host diet. Nearly all of the diet samples could be clustered by diet arm ($P < 0.003$, Fisher's exact test; Fig. 3f), despite the pre-existing inter-individual variation we observed during the baseline diets (Extended Data Fig. 6a, b). Still, subjects maintained their inter-individual differences on a taxonomic level on the diet arms (Extended Data Fig. 6c). Of the three RNA-seq samples on the animal-based diet that clustered with samples from the plant-based diet, all were taken on day 3 of the diet arm. In contrast, all RNA-seq samples from the final day of the diet arms (day 4) clustered by diet (Fig. 3f).

Remarkably, the plant- and animal-based diets also elicited transcriptional responses that were consistent with known differences in gene abundance between the gut microbiomes of herbivorous and carnivorous mammals, such as the trade-offs between amino acid catabolism versus biosynthesis, and in the interconversions of phosphoenolpyruvate (PEP) and oxaloacetate[2] (Fig. 3g, h). The former pathway favours amino acid catabolism when protein is abundant[2], and we speculate that the latter pathway produces PEP for aromatic amino acid synthesis when protein is scarce[20]. In all 14 steps of these pathways, we observed fold changes in gene expression on the plant- and animal-based diets, the directions of which agreed with the previously reported differences between herbivores and carnivores ($P < 0.001$, Binomial test). Notably, this perfect agreement is not observed when the plant- and animal-based diets are only compared with their respective baseline periods, indicating that the expression patterns in Fig. 3g, h reflect functional changes from both diet arms (Supplementary Table 17).

Our findings that the human gut microbiome can rapidly switch between herbivorous and carnivorous functional profiles may reflect past selective pressures during human evolution. Consumption of animal foods by our ancestors was probably volatile, depending on season and stochastic foraging success, with readily available plant foods offering a fall-back source of calories and nutrients[21]. Microbial communities that could quickly, and appropriately, shift their functional repertoire in response to diet change would have subsequently enhanced human dietary flexibility. Examples of this flexibility may persist today in the form of the wide diversity of modern human diets[11].

We next examined whether, in addition to affecting the resident gut microbiota, either diet arm introduced foreign microorganisms into the distal gut. We identified foodborne bacteria on both diets using 16S rRNA gene sequencing. The cheese and cured meats included in the animal-based diet were dominated by lactic acid bacteria commonly used as starter cultures for fermented foods[22,23]: *Lactococcus lactis*, *Pediococcus acidilactici* and *Streptococcus thermophilus* (Fig. 4a). Common non-lactic-acid bacteria included several *Staphylococcus* taxa; strains from this genus are often used when making fermented sausages[23]. During the animal-based diet, three of the bacteria associated with cheese and cured meats (*L. lactis*, *P. acidilactici* and *Staphylococcus*) became significantly more prevalent in faecal samples ($P < 0.05$, Wilcoxon signed-rank test; Extended Data Fig. 7c), indicating that bacteria found in common fermented foods can reach the gut at abundances above the detection limit of our sequencing experiments (on average 1 in 4×10^4 gut bacteria; Supplementary Table 6).

We also sequenced the internal transcribed spacer (ITS) region of the rRNA operon from community DNA extracted from food and

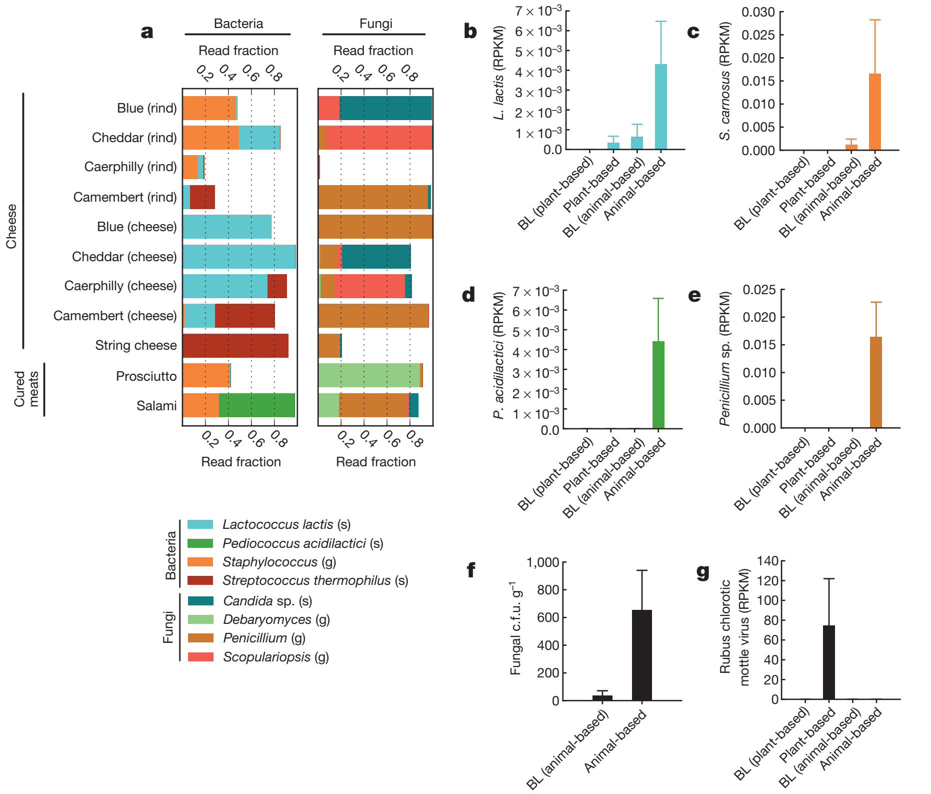


Figure 4 | Foodborne microbes are detectable in the distal gut. **a**, Common bacteria and fungi associated with the animal-based diet menu items, as measured by 16S rRNA and ITS gene sequencing, respectively. Taxa are identified on the genus (g) and species (s) level. A full list of foodborne fungi and bacteria on the animal-based diet can be found in Supplementary Table 21. Foods on the plant-based diet were dominated by matches to the Streptophyta, which derive from chloroplasts within plant matter (Extended Data Fig. 7a). **b–e**, Faecal RNA transcripts were significantly enriched ($q < 0.1$, Kruskal–Wallis test; $n = 6$–10 samples per diet arm) for several food-associated microbes on the animal-based diet relative to baseline (BL) periods, including *Lactococcus lactis* (**b**), *Staphylococcus carnosus* (**c**), *Pediococcus acidilactici* (**d**) and a *Penicillium* sp. (**e**). A complete table of taxa with significant expression differences can be found in Supplementary Table 22. **f**, Fungal concentrations in faeces before and 1–2 days after the animal-based diet were also measured using culture media selective for fungal growth (plate count agar with milk, salt and chloramphenicol). Post-diet faecal samples exhibit significantly higher fungal concentrations than baseline samples ($P < 0.02$; two-sided Mann–Whitney U test; $n = 7$–10 samples per diet arm). c.f.u., colony-forming units. **g**, Rubus chlorotic mottle virus transcripts increase on the plant-based diet ($q < 0.1$, Kruskal–Wallis test; $n = 6$–10 samples per diet arm). **b–g**, Bar charts all display mean ± s.e.m.

faecal samples to study the relationship between diet and enteric fungi, which so far remains poorly characterized (Supplementary Table 18). Menu items on both diets were colonized by the genera *Candida*, *Debaryomyces*, *Penicillium* and *Scopulariopsis* (Fig. 4a and Extended Data Fig. 7a), which are often found in fermented foods[22]. A *Penicillium* sp. and *Candida* sp. were consumed in sufficient quantities on the animal- and plant-based diets to show significant ITS sequence increases on those respective diet arms (Extended Data Fig. 7b, c).

Microbial culturing and re-analysis of our RNA-seq data suggested that foodborne microbes survived transit through the digestive system and may have been metabolically active in the gut. Mapping RNA-seq reads to an expanded reference set of 4,688 genomes (see Methods) revealed a significant increase on the animal-based diet of transcripts expressed by food-associated bacteria (Fig. 4b–d) and fungi (Fig. 4e; $q < 0.1$, Kruskal–Wallis test). Many dairy-associated microbes remained viable after passing through the digestive tract, as we isolated 19 bacterial and fungal strains with high genetic similarity (>97% ITS or 16S rRNA) to microbes cultured from cheeses fed to the subjects (Supplementary Table 19). Moreover, *L. lactis* was more abundant in faecal cultures sampled after the animal-based diet, relative to samples from the preceding baseline period ($P < 0.1$; Wilcoxon signed-rank test). We also detected an overall increase in the faecal concentration of viable fungi on the animal-based diet (Fig. 4f; $P < 0.02$; Mann–Whitney U test). Interestingly, we detected RNA transcripts from multiple plant viruses (Extended Data Fig. 8). One plant pathogen, Rubus chlorotic

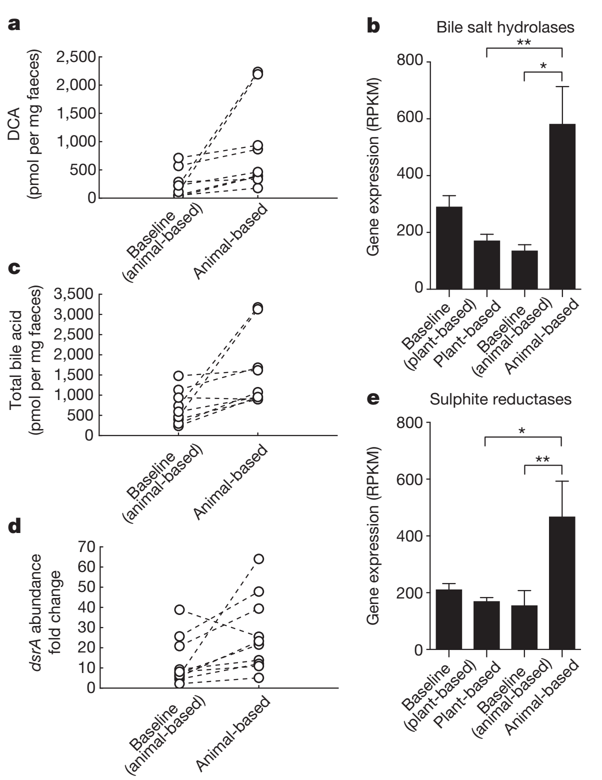


Figure 5 | Changes in the faecal concentration of bile acids and biomarkers for *Bilophila* on the animal-based diet. **a**, DCA, a secondary bile acid known to promote DNA damage and hepatic carcinomas[26], accumulates significantly on the animal-based diet ($P < 0.01$, two-sided Wilcoxon signed-rank test; see Supplementary Table 23 for the diet response of other secondary bile acids). **b**, RNA-seq data also supports increased microbial metabolism of bile acids on the animal-based diet, as we observed significantly increased expression of microbial bile salt hydrolases (K01442) during that diet arm (*$q < 0.05$, **$q < 0.01$, Kruskal–Wallis test; normalized to RPKM; $n = 8$–21 samples per diet arm). **c**, Total faecal bile acid concentrations also increase significantly on the animal-based diet, relative to the preceding baseline period ($P < 0.05$, two-sided Wilcoxon signed-rank test), but do not change on the plant-based diet (Extended Data Fig. 9). Bile acids have been shown to cause inflammatory bowel disease in mice by stimulating the growth of the bacterium *Bilophila*[6], which is known to reduce sulphite to hydrogen sulphide via the sulphite reductase enzyme DsrA (Extended Data Fig. 10). **d**, **e**, Quantitative polymerase chain reaction (PCR) showed a significant increase in microbial DNA coding for *dsrA* on the animal-based diet ($P < 0.05$; two-sided Wilcoxon signed-rank test) (**d**), and RNA-seq identified a significant increase in sulphite reductase expression (*$q < 0.05$, **$q < 0.01$, Kruskal–Wallis test; $n = 8$-21 samples/diet arm) (**e**). **b**, **e**, Bar graphs display mean ± s.e.m.

mottle virus, was only detectable on the plant-based diet (Fig. 4g). This virus infects spinach[24], which was a key ingredient in the prepared meals on the plant-based diet. These data support the hypothesis that plant pathogens can reach the human gut via consumed plant matter[25].

Finally, we found that microbiota changes on the animal-based diet could be linked to altered faecal bile acid profiles and the potential for human enteric disease. Recent mouse experiments have shown that high-fat diets lead to increased enteric deoxycholic concentrations (DCA); this secondary bile acid is the product of microbial metabolism and promotes liver cancer[26]. In our study, the animal-based diet significantly increased the levels of faecal DCA (Fig. 5a). Expression of bacterial genes encoding bile salt hydrolases, which are prerequisites for gut microbial production of DCA[27], was also significantly higher on the animal-based diet (Fig. 5b). Elevated DCA levels, in turn, may have contributed to the microbial disturbances on the animal-based diet, as this bile acid can inhibit the growth of members of the Bacteroidetes and Firmicutes phyla[28].

Mouse models have also provided evidence that inflammatory bowel disease can be caused by *B. wadsworthia*, a sulphite-reducing bacterium whose production of H_2S is thought to inflame intestinal tissue[6]. Growth of *B. wadsworthia* is stimulated in mice by select bile acids secreted while consuming saturated fats from milk. Our study provides several lines of evidence confirming that *B. wadsworthia* growth in humans can also be promoted by a high-fat diet. First, we observed *B. wadsworthia* to be a major component of the bacterial cluster that increased most while on the animal-based diet (cluster 28; Fig. 2 and Supplementary Table 8). This *Bilophila*-containing cluster also showed significant positive correlations with both long-term dairy ($P < 0.05$; Spearman correlation) and baseline saturated fat intake (Supplementary Table 20), supporting the proposed link to milk-associated saturated fats[6]. Second, the animal-based diet led to significantly increased faecal bile acid concentrations (Fig. 5c and Extended Data Fig. 9). Third, we observed significant increases in the abundance of microbial DNA and RNA encoding sulphite reductases on the animal-based diet (Fig. 5d, e). Together, these findings are consistent with the hypothesis that diet-induced changes to the gut microbiota may contribute to the development of inflammatory bowel disease. More broadly, our results emphasize that a more comprehensive understanding of diet-related diseases will benefit from elucidating links between nutritional, biliary and microbial dynamics.

METHODS SUMMARY

All experiments were performed under the guidance of the Harvard Committee on the Use of Human Subjects in Research; informed consent was obtained from all 11 subjects. Nine of those subjects participated in both diet arms, which were separated by 1 month. Each day, subjects logged their food intake and non-invasively sampled their gut microbiota. Each sample was either frozen immediately at −80 °C or briefly stored in personal −20 °C freezers before transport to the laboratory. Menu items for the animal- and plant-based diets were purchased at grocery stores and a restaurant, prepared by the experimenters, and distributed to the subjects daily. DNA was extracted from all faecal samples as previously described[29], sequenced using 16S rRNA- and ITS-specific primers, and analysed with the Quantitative Insights Into Microbial Ecology (QIIME) software package[30] and custom Python scripts. Three baseline days and 2 days on each experimental diet were selected for RNA-seq analysis[29]. Each RNA-seq data set was mapped to a functional database comprising 539 human-associated microbial genomes and a taxonomic identification database comprising 4,688 eukaryotic, prokaryotic and viral genomes. SCFA analysis was performed by gas chromatography, and bile-acid analysis used enzymatic assays and mass spectrometry.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 April; accepted 29 October 2013.
Published online 11 December 2013.

1. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–108 (2011).
2. Muegge, B. D. *et al.* Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332,** 970–974 (2011).
3. Duncan, S. H. *et al.* Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Appl. Environ. Microbiol.* **73,** 1073–1078 (2007).
4. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
5. Walker, A. W. *et al.* Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.* **5,** 220–230 (2011).
6. Devkota, S. *et al.* Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in *Il10*$^{-/-}$ mice. *Nature* **487,** 104–108 (2012).
7. Turnbaugh, P. J. *et al.* The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1,** 6ra14 (2009).
8. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
9. Faith, J. J., McNulty, N. P., Rey, F. E. & Gordon, J. I. Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science* **333,** 101–104 (2011).
10. Russell, W. R. *et al.* High-protein, reduced-carbohydrate weight-loss diets promote metabolite profiles likely to be detrimental to colonic health. *Am. J. Clin. Nutr.* **93,** 1062–1072 (2011).
11. Cordain, L. *et al.* Plant-animal subsistence ratios and macronutrient energy estimations in worldwide hunter-gatherer diets. *Am. J. Clin. Nutr.* **71,** 682–692 (2000).
12. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
13. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl Acad. Sci. USA* **107,** 14691–14696 (2010).
14. Reddy, B. S. Diet and excretion of bile acids. *Cancer Res.* **41,** 3766–3768 (1981).
15. Smith, E. A. & Macfarlane, G. T. Enumeration of amino acid fermenting bacteria in the human large intestine: effects of pH and starch on peptide metabolism and dissimilation of amino acids. *FEMS Microbiol. Ecol.* **25,** 355–368 (1998).
16. Smith, E. A. & Macfarlane, G. T. Enumeration of human colonic bacteria producing phenolic and indolic compounds: effects of pH, carbohydrate availability and retention time on dissimilatory aromatic amino acid metabolism. *J. Appl. Bacteriol.* **81,** 288–302 (1996).
17. Sinha, R. *et al.* High concentrations of the carcinogen 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP) occur in chicken but are dependent on the cooking method. *Cancer Res.* **55,** 4516–4519 (1995).
18. Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnol.* **31,** 814–821 (2013).
19. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
20. Pittard, J. & Wallace, B. J. Distribution and function of genes concerned with aromatic biosynthesis in *Escherichia coli. J. Bacteriol.* **91,** 1494–1508 (1966).
21. Hawkes, K., O'Connell, J. F. & Jones, N. G. Hunting income patterns among the Hadza: big game, common goods, foraging goals and the evolution of the human diet. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **334,** 243–250 (1991).
22. Bourdichon, F., Berger, B. & Casaregola, S. Safety demonstration of microbial food cultures (MFC) in fermented food products. *Bull. Int. Dairy Fed.* **455,** 1–66 (2012).
23. Nychas, G. J. & Arkoudelos, J. S. Staphylococci: their role in fermented sausages. *Soc. Appl. Bacteriol. Symp. Ser.* **19,** 167S–188S (1990).
24. McGavin, W. J. & Macfarlane, S. A. Rubus chlorotic mottle virus, a new sobemovirus infecting raspberry and bramble. *Virus Res.* **139,** 10–13 (2009).
25. Zhang, T. *et al.* RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* **4,** e3 (2006).
26. Yoshimoto, S. *et al.* Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. *Nature* **499,** 97–101 (2013).
27. Ridlon, J. M., Kang, D. J. & Hylemon, P. B. Bile salt biotransformations by human intestinal bacteria. *J. Lipid Res.* **47,** 241–259 (2006).
28. Islam, K. B. *et al.* Bile acid is a host factor that regulates the composition of the cecal microbiota in rats. *Gastroenterology* **141,** 1773–1781 (2011).
29. Maurice, C. F., Haiser, H. J. & Turnbaugh, P. J. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell* **152,** 39–50 (2013).
30. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335–336 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank A. Murray, G. Guidotti, E. O'Shea, J. Moffitt and B. Stern for insightful comments; M. Delaney for biochemical analyses; C. Daly, M. Clamp and C. Reardon for sequencing support; N. Fierer for providing ITS primers; A. Luong and K. Bauer for technical assistance; J. Brulc and R. Menon for nutritional guidelines; A. Rahman for menu suggestions; A. Must and J. Queenan for nutritional analysis; and our diet study volunteers for their participation. This work was supported by the National Institutes of Health (P50 GM068763), the Boston Nutrition Obesity Research Center (DK0046200), and the General Mills Bell Institute of Health and Nutrition.

Author Contributions L.A.D., R.J.D. and P.J.T. designed the study, and developed and prepared the diets. L.A.D., C.F.M., R.N.C., D.B.G., J.E.B., B.E.W. and P.J.T. performed the experimental work. A.V.L., A.S.D., Y.V., M.A.F. and S.B.B. conducted bile acid analyses. L.A.D. and P.J.T. performed computational analyses. L.A.D. and P.J.T. prepared the manuscript.

Author Information RNA-seq data have been deposited in the Gene Expression Omnibus under accession GSE46761; 16S and ITS rRNA gene sequencing reads have been deposited in MG-RAST under accession 6248. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.J.T. (pturnbaugh@fas.harvard.edu).

METHODS

Sample collection. All experiments were performed under the guidance of the Harvard Committee on the Use of Human Subjects in Research. We recruited 11 unrelated subjects ($n = 10$ per diet; 9 individuals completed both arms of the study). One participant suffered from a chronic gastrointestinal disease, but all other volunteers were otherwise healthy. The volunteers' normal bowel frequencies ranged from three times a day to once every other day. Three participants had taken antibiotics in the past year. Additional subject information is provided in Supplementary Table 2. Gut microbial communities were sampled and analysed from faeces[29,30]. Subjects were instructed to collect no more than one sample per day, but to log all bowel movements. No microbiota patterns were observed as a function of sampling time of day (data not shown). Subjects collected samples by placing disposable commode specimen containers (Claflin Medical Equipment) under their toilet seats before bowel movements. CultureSwabs (BD) were then used to collect faecal specimens for sequencing analyses, and larger collection tubes were provided for harvesting larger, intact stool samples (~10 g) for metabolic analyses. Each sample was either frozen immediately at −80 °C or briefly stored in personal −20 °C freezers before transport to the laboratory.

Diet design. We constructed two diet arms, each of which consisted mostly of plant- or animal-based foods (Extended Data Fig. 1). Subjects on the plant-based diet ate cereal for breakfast and precooked meals made of vegetables, rice and lentils for lunch and dinner (see Supplementary Table 1 for a full list of diet ingredients). Fresh and dried fruits were provided as snacks on this diet. Subjects on the animal-based diet ate eggs and bacon for breakfast, and cooked pork and beef for lunch. Dinner consisted of cured meats and a selection of four cheeses. Snacks on this diet included pork rinds, cheese and salami. Ingredients for the plant-based diet, dinner meats and cheeses for the animal-based diet, and snacks for both diets were purchased from grocery stores. Lunchmeats for the animal-based diet were prepared by a restaurant that was instructed to not add sauce to the food. On each diet arm, subjects were instructed to eat only provided foods or allowable beverages (water or unsweetened tea for both diets; coffee was allowed on the animal-based diet). They were also allowed to add one salt packet per meal, if desired for taste. Subjects could eat unlimited amounts of the provided foods. Outside of the 5-day diet arms, subjects were instructed to eat normally.

Food logs, subject metadata and dietary questionnaires. Subjects were given notepads to log their diet, health and bowel movements during the study. Subjects transcribed their notepads into digital spreadsheets when the study ended. Each ingested food (including foods on the diet arm) was recorded, as well as data on time, location, portion size, and food brand. Subjects were provided with pocket digital scales (American Weigh) and a visual serving size guide to aid with quantifying the amount of food consumed. Each day, subjects tracked their weight using either a scale provided in the lab, or their own personal scales at home. While on the animal-based diet, subjects were requested to measure their urinary ketone levels using provided Ketostix strips (Bayer; Extended Data Fig. 1). If subjects recorded a range of ketone levels (the Ketostix colour key uses a range-based reporting system) the middle value of that range was used for further analysis. Subjects were encouraged to record any discomfort they experienced while on either diet (for example, bloating, constipation). Subjects tracked all bowel movements, regardless of whether or not they collected samples, recording movement time, date and location, and qualitatively documented stool colour, odour and type[31]. Subjects were also asked to report when they observed stool staining from food dyes consumed at the beginning and end of each diet arm (Extended Data Fig. 3a).

Diet quantification. We quantified subjects' daily nutritional intake during the study using CalorieKing and Nutrition Data System for Research (NDSR). The CalorieKing food database was accessed via the CalorieKing Nutrition and Exercise Manager software (version 4.1.0). Subjects' food items were manually transferred from digital spreadsheets into the CalorieKing software, which then tabulated each food's nutritional content. Macronutrient content per serving was calculated for each of the prepared meals on the animal- and plant-based diet using lists of those meals' ingredients. Nutritional data was outputted from CalorieKing in CSV format and parsed for further analysis using a custom Python script. NDSR intake data were collected and analysed using Nutrition Data System for Research software version 2012, developed by the Nutrition Coordinating Center (NCC), University of Minnesota. We estimated subjects' long-term diet using the National Cancer Institute's Diet History Questionnaire II (DHQ)[32]. We used the DHQ to quantify subjects' annual diet intake, decomposed into 176 nutritional categories. Subjects completed the yearly, serving-size-included version of the DHQ online using their personal computers. We parsed the survey's results using Diet*Calc software (version 1.5; Risk Factor Monitoring and Methods Branch, National Cancer Institute) and its supplied 'Food and Nutrient Database' and 'dhqweb.yearly.withserv.2010.qdd' QDD file. There was good agreement between subjects' diets as measured by CalorieKing, the NDSR, and the DHQ: 18 of 20 nutritional comparisons between pairs of databases showed significant correlations (Supplementary Table 3). Unless specified, nutritional data presented in this manuscript reflect CalorieKing measurements.

16S rRNA gene sequencing and processing. Temporal patterns of microbial community structure were analysed from daily faecal samples collected across each diet (Extended Data Fig. 1). Samples were kept at −80 °C until DNA extraction with the PowerSoil bacterial DNA extraction kit (MoBio). The V4 region of the 16S rRNA gene was PCR amplified in triplicate, and the resulting amplicons were cleaned, quantified and sequenced on the Illumina HiSeq platform according to published protocols[33,34] and using custom barcoded primers (Supplementary Table 6). Raw sequences were processed using the QIIME software package[30]. Only full-length, high-quality reads ($-r = 0$) were used for analysis. OTUs were picked at 97% similarity against the Greengenes database[35] (constructed by the nested_gg_workflow.py QiimeUtils script on 4 February, 2011), which we trimmed to span only the 16S rRNA region flanked by our sequencing primers (positions 521–773). In total, we characterized an average of 43,589 ± 1,826 16S rRNA sequences for 235 samples (an average of 0.78 samples per person per study day; Supplementary Table 6). Most of the subsequent analysis of 16S rRNA data, including calculations of α and β diversity, were performed using custom Python scripts, the SciPy Python library[36], and the Pandas Data Analysis Library[37]. Correction for multiple hypothesis testing used the fdrtool[38] R library, except in the case of small test numbers, in which case the Bonferroni correction was used.

OTU clustering. We used clustering to simplify the dynamics of thousands of OTUs into a limited number of variables that could be more easily visualized and manually inspected. Clustering was performed on normalized OTU abundances. Such abundances are traditionally computed by scaling each sample's reads to sum to a fixed value (for example, unity); this technique is intended to account for varying sequencing depth between samples. However, this standard technique may cause false relationships to be inferred between microbial taxa, as increases in the abundance of one microbial group will cause decreases in the fractional abundance of other microbes (this artefact is known as a 'compositional' effect[39]). To avoid compositional biases, we used an alternative normalization approach, which instead assumes that no more than half of the OTUs held in common between two microbial samples change in abundance. This method uses a robust (outlier-resistant) regression to estimate the median OTU fold change between communities, by which it subsequently rescales all OTUs. To simplify community dynamics further, we only included in our clustering model OTUs that comprised 95% of total reads (after ranking by normalized abundance).

Abundances for each included OTU were then converted to log space and median centred. We computed OTU pairwise distances using the Pearson correlation (OTU abundances across all subjects and time points were used). The resulting distance matrix was subsequently inputted into Scipy's hierarchical clustering function ('fcluster').

Default parameters were used for fcluster, with the exception of the clustering criterion, which was set to 'distance', and the clustering threshold, which was set to '0.7'. These parameters were selected manually so that cluster boundaries visually agreed with the correlation patterns plotted in a matrix of pairwise OTU distances. Statistics on cluster abundance during baseline and diet periods were computed by taking median values across date ranges. Baseline date ranges were the 4 days preceding each diet arm (that is, days −4 to −1). Date ranges for the diet arms were chosen so as to capture the full effects of each diet. These ranges were not expected to perfectly overlap with the diet arms themselves, due to the effects of diet transit time. We therefore chose diet arm date ranges that accounted for transit time (as measured by food dye; Extended Data Fig. 3a), picking ranges that began 1 day after foods reached the gut, and ended 1 day before the last diet arm meal reached the gut. These criteria led microbial abundance measurements on the plant-based diet to span days 2–4 of that study arm, and animal-based diet measurements to span days 2–5 of that diet arm.

RNA-seq sample preparation and sequencing. We measured community-wide gene expression using meta-transcriptomics[7,29,40,41] (Supplementary Table 12). Samples were selected on the basis of our prior 16S rRNA gene-sequencing-based analysis, representing 3 baseline days and 2 time points on each diet (n = 5–10 samples per time point; Extended Data Fig. 1). Microbial cells were lysed by a bead beater (BioSpec Products), total RNA was extracted with phenol:chloroform:isoamyl alcohol (pH 4.5, 125:24:1, Ambion 9720) and purified using Ambion MEGAClear columns (Life Technologies), and rRNA was depleted via Ambion MICROBExpress subtractive hybridization (Life Technologies) and custom depletion oligonucleotides. The presence of genomic DNA contamination was assessed by PCR with universal 16S rRNA gene primers. cDNA was synthesized using SuperScript II and random hexamers ordered from Invitrogen (Life Technologies), followed by second-strand synthesis with RNaseH and *E.coli* DNA polymerase (New England Biolabs). Samples were prepared for sequencing with an Illumina HiSeq instrument after enzymatic fragmentation (NEBE6040L/M0348S). Libraries were quantified by quantitative reverse transcriptase PCR (qRT–PCR) according to the Illumina

protocol. qRT–PCR assays were run using ABsoluteTM QPCR SYBR Green ROX Mix (Thermo Scientific) on a Mx3000P QPCR System instrument (Stratagene). The size distribution of each library was quantified on an Agilent HS-DNA chip. Libraries were sequenced using the Illumina HiSeq platform.

Functional analysis of RNA-seq data. We used a custom reference database of bacterial genomes to perform functional analysis of the RNA-seq data[29]. This reference included 538 draft and finished bacterial genomes obtained from human-associated microbial isolates[42], and the *Eggerthella lenta* DSM22 (ref. 7) reference genome. All predicted proteins from the reference genome database were annotated with KEGG[19] orthologous groups (KOs) using the KEGG database (version 52; BLASTX *e* value $< 10^{-5}$, bit score >50, and $>50\%$ identity). For query genes with multiple matches, the annotated reference gene with the lowest *e* value was used. When multiple annotated genes with an identical *e* value were encountered after a BLAST query, we included all KOs assigned to those genes. Genes from the database with significant homology (BLASTN *e* value $< 10^{-20}$) to non-coding transcripts from the 539 microbial genomes were excluded from subsequent analysis. High-quality reads (see Supplementary Table 12 for sequencing statistics) were mapped using SSAHA2 (ref. 43) to our reference bacterial database and the Illumina adaptor sequences (SSAHA2 parameters: '-best 1 -score 20 -solexa'). The number of transcripts assigned to each gene was then tallied and normalized to RPKM. To account for genes that were not detected owing to limited sequencing depth, a pseudocount of 0.01 was added to all samples. Samples were clustered in Matlab (version 7.10.0) using a Spearman distance matrix (commands: pdist, linkage, and dendrogram). Genes were grouped by taxa, genomes and KOs by calculating the cumulative RPKM for each sample. HUMAnN[44] was used for metabolic reconstruction from metagenomic data followed by LefSe[45] analysis to identify significant biomarkers. A modified version of the 'SConstruct' file was used to input KO counts into the HUMAnN pipeline for each RNA-seq data set. We then ran LefSe on the resulting KEGG module abundance file using the '-o 1000000' flag.

Taxonomic analysis of RNA-seq data. We used Bowtie 2 read alignment program[46] and the Integrated Microbial Genomes (IMG; version 3.5) database[47] to map RNA-seq reads to a comprehensive reference survey of prokaryotic, eukaryotic and viral genomes. Our reference survey included all 2,809 viral genomes in IMG (as of version 3.5), a set of 1,813 bacterial and archaeal genomes selected to minimize strain redundancy[48], and 66 genomes spanning the Eukarya except for the plants and non-nematode Bilateria. Reads were mapped to reference genomes using Bowtie 2, which was configured to analyse mated paired-end reads, and return fragments with a minimum length of 150 bp and a maximum length of 600 bp. All other parameters were left to their default values. The number of base pairs in the reference genome data set exceeded Bowtie's reference size limit, so we split the reference genomes into four subsets. Each read was mapped to each of these four sub-reference data sets, and the results were merged by picking the highest-scoring match across the sub-references. We settled tied scores by randomly choosing one of the best-scoring matches. To measure more precisely the presence or absence of specific taxa, we filtered out reads that mapped to more than one reference sequence. Raw read counts were computed for each reference genome by counting the number of reads that mapped to coding sequences according to the IMG annotations; these counts were subsequently normalized using RPKM scaling. Our analysis pipeline associated several sequences with marine algae, which are unlikely to colonize the human gut. We also detected a fungal pathogen exclusively in samples from subjects consuming the animal-based diet (*Neosartorya fischeri*); this taxon was suspected of being a misidentified cheese fungus, owing to its relatedness to *Penicillium*. We thus reanalysed protist and *N. fischeri* reads associated with potentially mis-annotated taxa using BLAST searches against the NCBI non-redundant database, and we assigned taxonomy manually based on the most common resulting hits (Extended Data Fig. 8).

qPCR. Community DNA was isolated with the PowerSoil bacterial DNA extraction kit (MoBio). To determine the presence of hydrogen consumers, PCR was performed on faecal DNA using the following primer sets: sulphite reductase[6] (dsrA), forward, 5′-CCAACATGCACGGYTCCA-3′, reverse, 5′-CGTCGAACTTGAACTTGAACTTGTAGG-3′; and sulphate reduction[49,50] (aps reductase), forward, 5′-TGGCAGATMATGATYMACGG-3′, reverse, 5′-GGGCCGTAACCGTCCTTGAA-3′. qPCR assays were run using ABsoluteTM QPCR SYBR Green ROX Mix (Thermo Scientific) on a Mx3000P QPCR System instrument (Stratagene). Fold changes were calculated relative to the 16S rRNA gene using the $2^{-\Delta\Delta Ct}$ method and the same primers used for 16S rRNA gene sequencing.

SCFA measurements. Faecal SCFA content was determined by gas chromatography. Chromatographic analysis was carried out using a Shimadzu GC14-A system with a flame ionization detector (FID) (Shimadzu Corp). Fused silica capillary columns 30 m × 0.25 mm coated with 0.25 μm film thickness were used (Nukol for the volatile acids and SPB-1000 for the nonvolatile acids (Supelco Analytical). Nitrogen was used as the carrier gas. The oven temperature was 170 °C and the FID and injection port was set to 225 °C. The injected sample volume was 2 μl and the run time for each analysis was 10 min. The chromatograms and data integration was carried out using a Shimadzu C-R5A Chromatopac. A volatile acid mix containing 10 mM of acetic, propionic, isobutyric, butyric, isovaleric, valeric, isocaproic, caproic and heptanoic acids was used (Matreya). A non-volatile acid mix containing 10 mM of pyruvic and lactic and 5 mM of oxalacetic, oxalic, methyl malonic, malonic, fumaric and succinic acids was used (Matreya). A standard stock solution containing 1% 2-methyl pentanoic acid (Sigma-Aldrich) was prepared as an internal standard control for the volatile acid extractions. A standard stock solution containing 50 mM benzoic acid (Sigma-Aldrich) was prepared as an internal standard control for the non-volatile acid extractions.

Samples were kept frozen at −80 °C until analysis. The samples were removed from the freezer and 1,200 μl of water was added to each thawed sample. The samples were vortexed for 1 min until the material was homogenized. The pH of the suspension was adjusted to 2–3 by adding 50 μl of 50% sulphuric acid. The acidified samples were kept at room temperature (∼21 °C) for 5 min and vortexed briefly every minute. The samples were centrifuged for 10 min at 5,000*g*. Five-hundred microlitres of the clear supernatant was transferred into two tubes for further processing. For the volatile extraction, 50 μl of the internal standard (1% 2-methyl pentanoic acid solution) and 500 μl of ethyl ether anhydrous were added. The tubes were vortexed for 30 s and then centrifuged at 5,000*g* for 10 min. One microlitre of the upper ether layer was injected into the chromatogram for analysis. For the nonvolatile extraction, 50 μl of the internal standard (50 mM benzoic acid solution) and 500 μl of boron trifluoride-methanol solution (Sigma-Aldrich) were added to each tube. These tubes were incubated overnight at room temperature. One millilitre of water and 500 μl of chloroform were added to each tube. The tubes were vortexed for 30 s and then centrifuged at 5,000*g* for 10 min. One microlitre of the lower chloroform layer was injected into the chromatogram for analysis. Five-hundred microlitres of each standard mix was used and the extracts prepared as described for the samples. The retention times and peak heights of the acids in the standard mix were used as references for the sample unknowns. These acids were identified by their specific retention times and the concentrations determined and expressed as mM concentrations per gram of sample.

Bulk bile acid quantification. Faecal bile acid concentration was measured as described previously[51]. One-hundred milligrams of lyophilized stool was heated to 195 °C in 1 ml of ethylene glycol KOH for 2 h, neutralized with 1 ml of saline and 0.2 ml of concentrated HCl, and extracted into 6 ml of diethyl ether three times. After evaporation of the ether, the sample residues were dissolved in 6 ml of methanol and subjected to enzymatic analysis. Enzymatic reaction mixtures consisted of 66.5 mmol l^{-1} Tris, 0.33 mmol l^{-1} EDTA, 0.33 mol l^{-1} hydrazine hydrate, 0.77 mmol l^{-1} NAD (N 7004, Sigma-Aldrich), 0.033 U ml^{-1} 3α-hydroxysteroid dehydrogenase (Sigma-Aldrich) and either sample or standard (taurocholic acid; Sigma-Aldrich) dissolved in methanol. After 90 min of incubation at 37 °C, absorbance was measured at 340 nm.

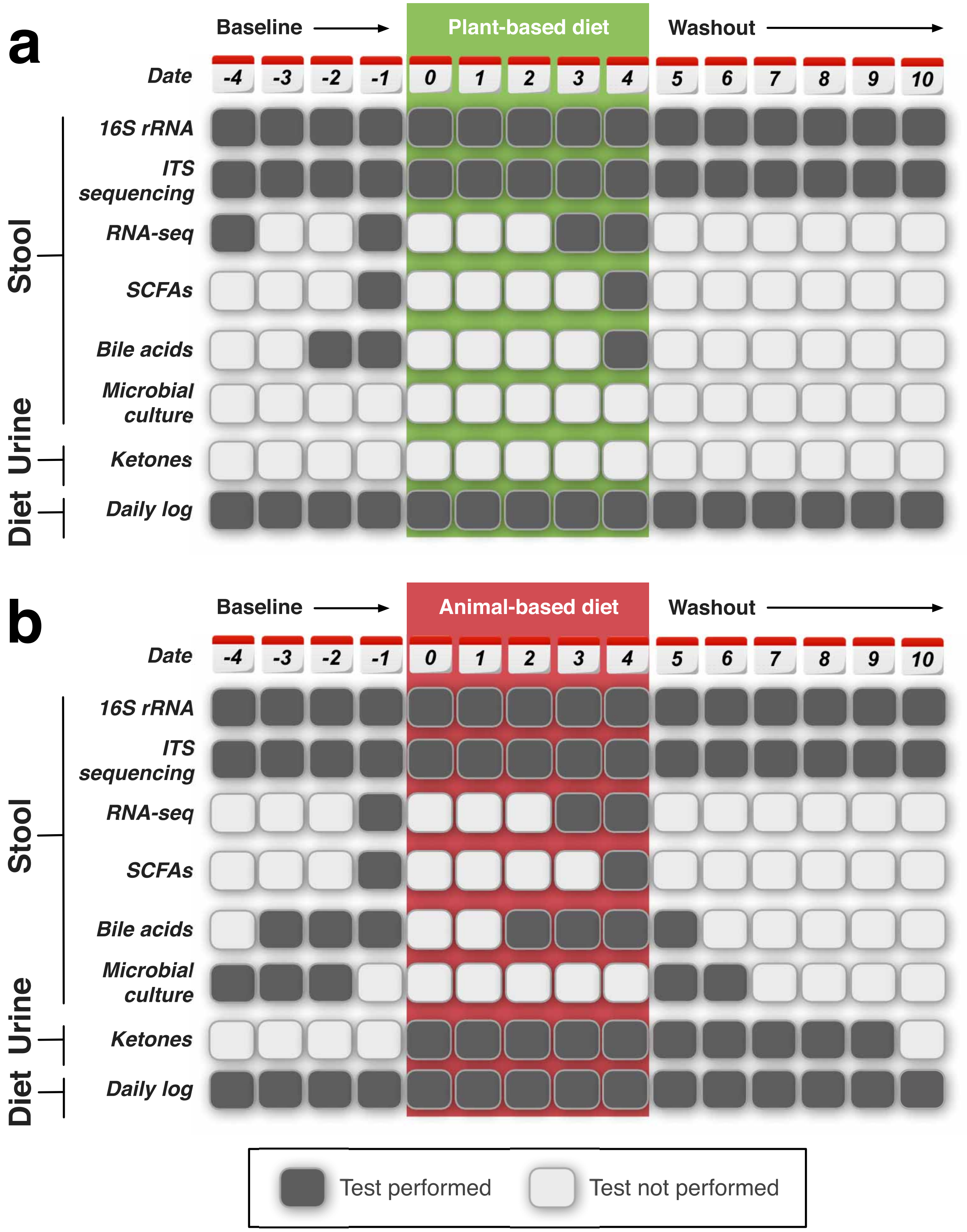
Measurement of primary and secondary bile acids. Profiling of faecal primary and secondary bile acids was performed using a modified version of a method described previously[52]. To a suspension of ∼100 mg of stool and 0.25 ml of water in a 4 ml Teflon-capped glass vial was added 200 mg of glass beads. The suspension was homogenized by vortexing for 60–90 s. Ethanol (1.8 ml) was added, and the suspension was heated with stirring in a heating block at 80 °C for 1.5 h. The sample was cooled, transferred to a 2 ml Eppendorf tube, and centrifuged at 12,225*g* for 1–2 min. The supernatant was removed and retained. The pellet was resuspended in 1.8 ml of 80% aqueous ethanol, transferred to the original vial, and heated to 80 °C for 1.5 h. The sample was centrifuged again, and the supernatant was removed and added to the first extraction supernatant. The pellet was resuspended in 1.8 ml of chloroform:methanol (1:1 v/v) and refluxed for 30–60 min. The sample was centrifuged, and the supernatant removed and concentrated to dryness on a rotary evaporator. The ethanolic supernatants were added to the same flask, the pH was adjusted to neutrality by adding aqueous 0.01N HCl, and the combined extracts were evaporated to dryness. The dried extract was resuspended in 1 ml of 0.01N aqueous HCl by sonication for 30 min. A BIO-RAD Poly-Prep chromatography column (0.8 × 4 cm) was loaded with Lipidex 1000 as a slurry in MeOH, allowed to pack under gravity to a final volume of 1.1 ml, and washed with 10 ml of distilled water. The suspension was filtered through the bed of Lipidex 1000 and the effluent was discarded. The flask was washed with 3 × 1 ml of 0.01N HCl, the washings were passed through the gel, and the bed was washed with 4 ml of distilled water. Bile acids and sterols were recovered by elution of the Lipidex gel bed with 8 ml of methanol. A BIO-RAD Poly-Prep chromatography column (0.8 × 4 cm) was loaded with washed SP-Sephadex as a slurry in 72% aqueous MeOH to a final volume of 1.1 ml. The methanolic extract was passed through the SP-Sephadex column, and the column was washed with 4 ml of 72% aqueous methanol. The extract and wash were combined, and the pH was brought to neutral with 0.04N aqueous NaOH. A BIO-RAD Poly-Prep chromatography column (0.8 × 4 cm) was loaded with Lipidex-DEAP, prepared in the acetate form,

as a slurry in 72% aqueous MeOH to a final volume of 1.1 ml. The combined neutralized effluent was applied to the column, and the solution was eluted using air gas pressure (flow rate ~25 ml h^{-1}). The flask and column were washed with 2×2 ml of 72% aqueous ethanol, and the sample and washings were combined to give a fraction of neutral compounds including sterols. Unconjugated bile acids were eluted using 4 ml of 0.1 M acetic acid in 72% (v/v) aqueous ethanol that had been adjusted to pH 4.0 by addition of concentrated ammonium hydroxide. The fraction containing bile acids was concentrated to dryness on a rotary evaporator. The bile acids were converted to their corresponding methyl ester derivatives by the addition of 0.6 ml of MeOH followed by 40 μl of a 2.0 M solution of (trimethylsilyl)-diazomethane in diethyl ether. The solution was divided in half, and each half of the sample was concentrated to dryness on a rotary evaporator. The bile acids in the first half of the sample were converted to their corresponding trimethylsilyl ether derivatives by the addition of 35 μl of a 2:1 solution of *N*,*O*-bis(trimethylsilyl) trifluoroacetamide and chlorotrimethylsilane and analysed by gas chromatography-mass spectrometry (GC-MS). The identities of individual bile acids were determined by comparison of retention time and fragmentation pattern to known standards. Both the ratio of cholest-3-ene to deoxycholic acid in the sample and the amount of internal standard to be added were determined by integrating peak areas. A known amount of the internal standard, 5- cholestane-3-ol (5-coprostanol), was added to the second half of the sample (0.003–0.07 mmol). The bile acids in the second half of the sample were converted to their corresponding trimethylsilyl ether derivatives by the addition of 35 μl of a 2:1 solution of *N*,*O*-bis(trimethylsilyl)trifluoroacetamide and chlorotrimethylsilane and analysed by GC-MS. Amounts of individual bile acids were determined by dividing integrated bile acid peak area by the internal standard peak area, multiplying by the amount of internal standard added, and then dividing by half of the mass of faecal matter extracted. In the event that the first half of the sample contained cholest-3-ene, the coprostanol peak area in the second half of the sample was corrected by subtracting the area of the cholest-3-ene peak, determined by applying the cholest-3-ene:deoxycholic acid ratio calculated from the first half of the sample.

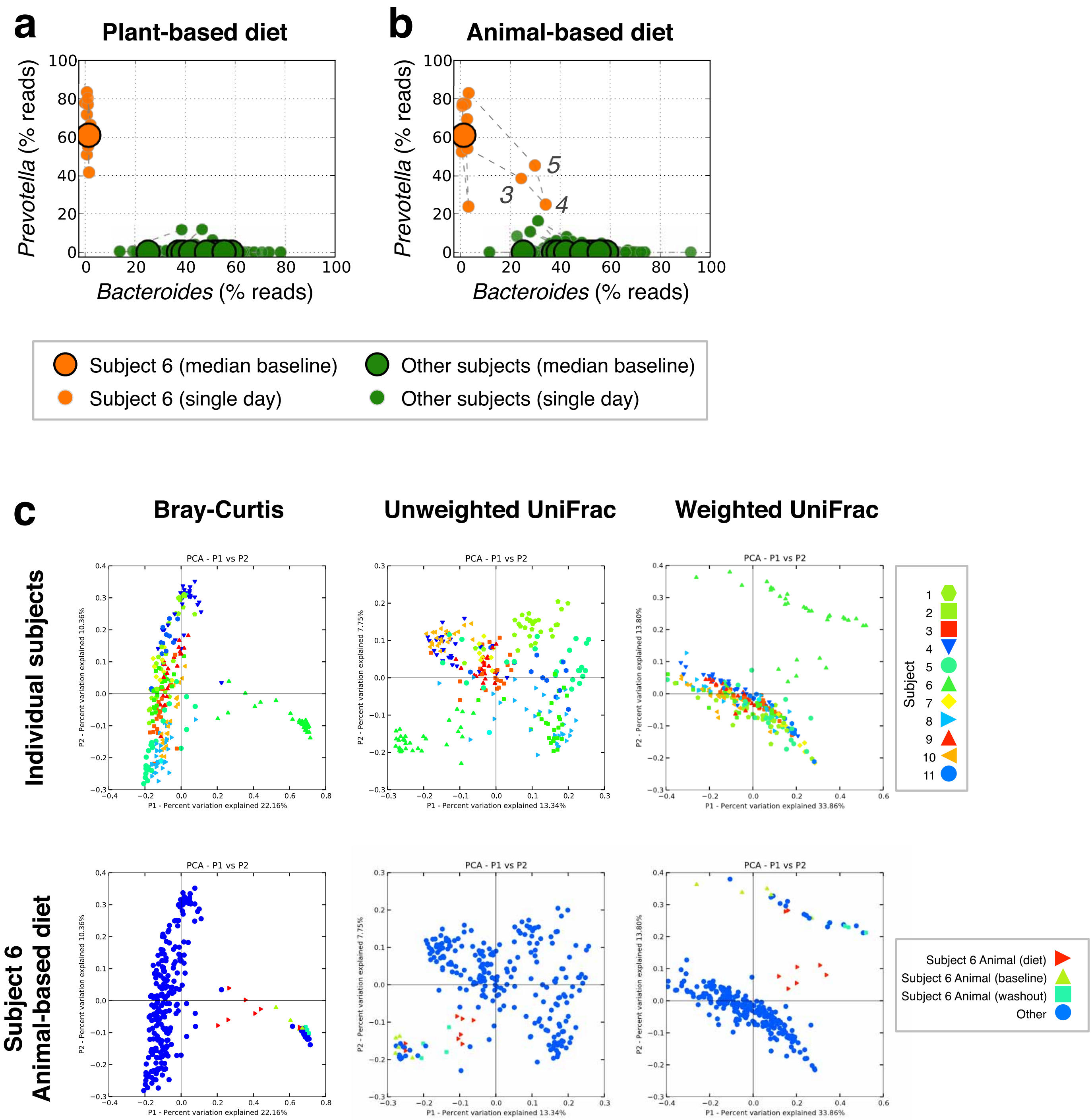
ITS sequencing. Fungal amplicon libraries were constructed with primers that target the ITS, a region of the nuclear ribosomal RNA cistron shown to promote successful identification across a broad range of fungal taxa[53]. We selected primers—ITS1f (ref. 54) and ITS2 (ref. 55)—focused on the ITS1 region because it provided the best discrimination between common cheese-associated fungi in preliminary *in silico* tests. Multiplex capability was achieved by adding Golay barcodes to the ITS2 primer. Owing to relatively low concentrations, fungal DNA was amplified in three serial PCR reactions, with the first reaction using 1 μl of the PowerSoil DNA extract, and the subsequent two reactions using 1 μl of the preceding PCR product as the template. In each round of PCR, sample reactions were performed in triplicate and then combined. Barcoded amplicons were cleaned, quantified and pooled to achieve approximately equal amounts of DNA from each sample using methods identical to those used for 16S. We gel purified the pool, targeting amplicons between 150 bp and 500 bp in size, and submitted it for Illumina sequencing. Preliminary taxonomic assignments of ITS reads using the 12_11 UNITE OTUs ITS database (see http://qiime.org) resulted in many unassigned reads. To improve the percentage of reads assigned, we created our own custom database of ITS1 sequences. We extracted ITS sequences from GenBank by targeting specific collections of reliable ITS sequences (for example, AFTOL, Fungal Barcoding Consortium) and by searching for sequences of yeasts and filamentous fungi that have been previously isolated from dairy and other food ecosystems. We also retrieved a wider range of fungi for our database by searching GenBank with the query "internal transcribed spacer[All Fields] AND fungi NOT 'uncultured'". Sequences that did not contain the full ITS1 were removed. We also included reference OTUs that were identified as widespread cheese fungi in a survey of cheese rinds (B.E.W., J.E.B. and R.J.D., unpublished observations), but were not in public databases.

Microbial culturing. Faecal samples were cultured under conditions permissive for the growth of food-derived microbes. Faecal samples were suspended in a volume of PBS equivalent to ten times their weight. Serial dilutions were prepared and plated on brain heart infusion agar (BD Biosciences) supplemented with 100 μg ml^{-1} cycloheximide, an antifungal agent, and plate count agar with milk and salt (per litre: 5 g tryptone, 2.5 g yeast extract, 1 g dextrose, 1 g whole milk powder, 30 g NaCl, 15 g agar) supplemented with 50 μg ml^{-1} chloramphenicol, an antibacterial agent. Plates were incubated under aerobic conditions at room temperature for 7 days. Plates supplemented with chloramphenicol that yielded significant growth of bacteria, as determined by colony morphology, were excluded from further analysis. Plates were examined by eye for bacterial colonies or fungal foci whose morphological characteristics were similar to previously characterized food-derived microbes. Candidate food-derived microbes were isolated and identified by Sanger sequencing of the 16S rRNA gene (for bacteria; primers used were 27f, 5-AGAGTTTGATCCTGGCTCAG and 1492r, 5-GGTTACCTTGTTACGACTT) or ITS region (for fungi; primers used were ITS1f, 5-CTTGGTCATTTAGAGGAA GTAA and ITS4, 5 TCCTCCGCTTATTGATATGC). After select colonies had been picked for isolation, the surface of each plate was scraped with a razor blade to collect all remaining colonies, and material was suspended in PBS. Dilutions were pooled, and DNA was extracted from the resulting pooled material using a PowerSoil kit (MoBio). The remaining pooled material was stocked in 20% glycerol and stored at −80 °C.

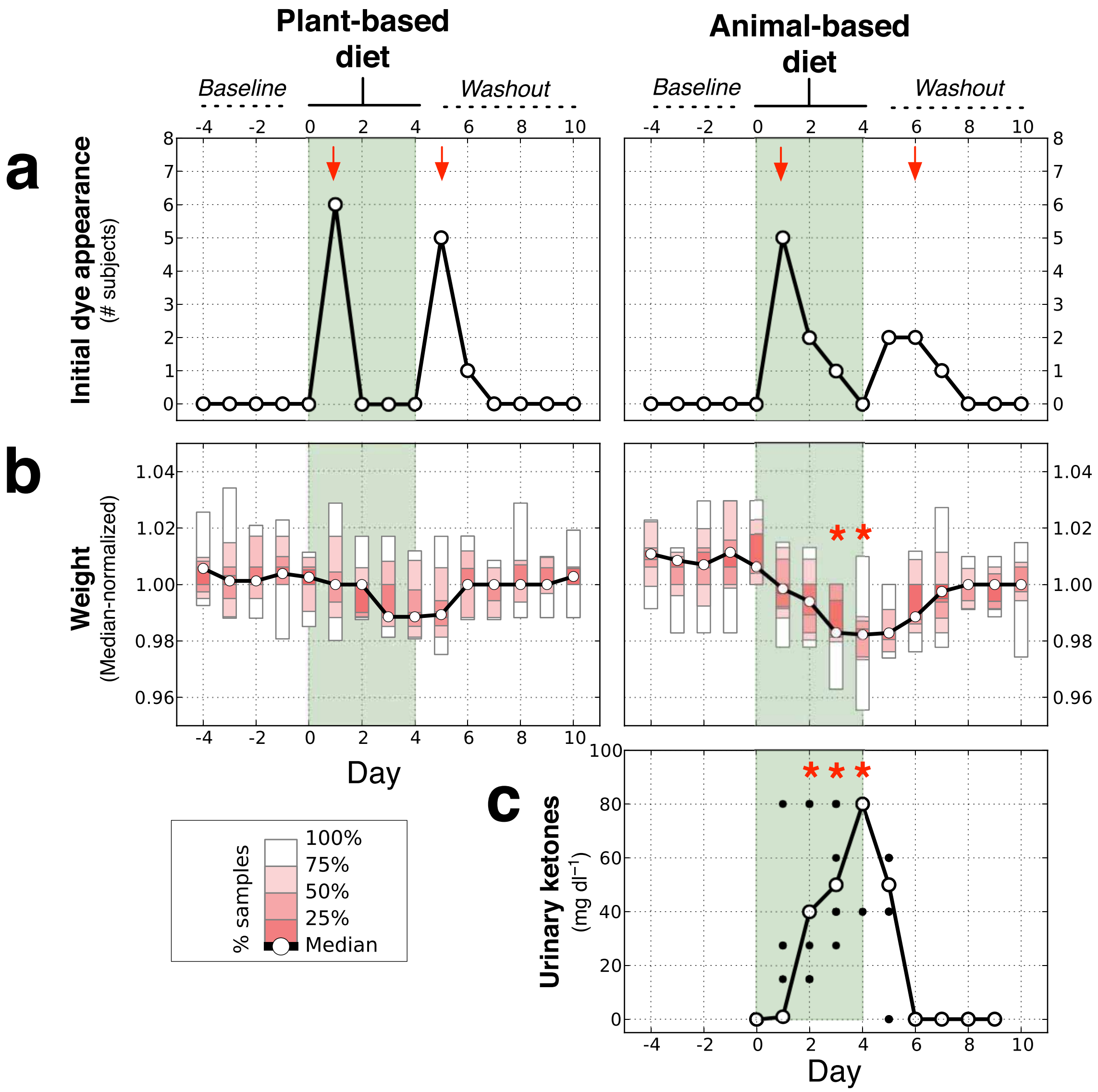
31. Lewis, S. J. & Heaton, K. W. Stool form scale as a useful guide to intestinal transit time. *Scand. J. Gastroenterol.* **32,** 920–924 (1997).
32. National Institutes of Health. *Diet History Questionnaire Version 2.0* (National Institutes of Health, Applied Research Program, National Cancer Institute, 2010).
33. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl Acad. Sci. USA* **108** (suppl. 1), 4516–4522 (2011).
34. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6,** 1621–1624 (2012).
35. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).
36. Jones, E. *et al.* SciPy: open source scientific tools for Python (2001).
37. McKinney, W. Data structures for statistical computing in Python. *Proc. 9th Python Sci. Conf.* 51–56 (2010).
38. Strimmer, K. fdrtool: a versatile R package for estimating local and tail areabased false discovery rates. *Bioinformatics* **24,** 1461–1462 (2008).
39. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8,** e1002687 (2012).
40. Turnbaugh, P. J. *et al.* Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc. Natl Acad. Sci. USA* **107,** 7503–7508 (2010).
41. Rey, F. E. *et al.* Dissecting the *in vivo* metabolic potential of two human gut acetogens. *J. Biol. Chem.* **285,** 22082–22090 (2010).
42. Nelson, K. E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328,** 994–999 (2010).
43. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
44. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLOS Comput. Biol.* **8,** e1002358 (2012).
45. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12,** R60 (2011).
46. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
47. Markowitz, V. M. *et al.* IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* **40,** D115–D122 (2012).
48. Martin, J. *et al.* Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PLoS ONE* **7,** e36427 (2012).
49. Deplancke, B. *et al.* Molecular ecological analysis of the succession and diversity of sulfate-reducing bacteria in the mouse gastrointestinal tract. *Appl. Environ. Microbiol.* **66,** 2166–2174 (2000).
50. Stewart, J. A., Chadwick, V. S. & Murray, A. Carriage, quantification, and predominance of methanogens and sulfate-reducing bacteria in faecal samples. *Lett. Appl. Microbiol.* **43,** 58–63 (2006).
51. Porter, J. L. *et al.* Accurate enzymatic measurement of fecal bile acids in patients with malabsorption. *J. Lab. Clin. Med.* **141,** 411–418 (2003).
52. Setchell, K. D., Lawson, A. M., Tanida, N. & Sjovall, J. General methods for the analysis of metabolic profiles of bile acids and related compounds in feces. *J. Lipid Res.* **24,** 1085–1100 (1983).
53. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for *Fungi. Proc. Natl Acad. Sci. USA* **109,** 6241–6246 (2012).
54. Gardes, M. & Bruns, T. D. ITS primers with enhanced specificity for basidiomycetes application to the identification of mycorrhizae and rusts. *Mol. Ecol.* **2,** 113–118 (1993).
55. White, T. J., Bruns, T., Lee, S. & Taylor, J. in *PCR Protocols: A Guide to Methods and Applications* (eds Gelfand, D. H., Innis, M. A., Shinsky, J. J. & White, T. J.) 315–322 (1990).
56. Walker, H. K., Hall, W. D., Hurst, J. W., Comstock, J. P. & Garber, A. J. *Ketonuria* 3rd edn (Butterworths, 1990).



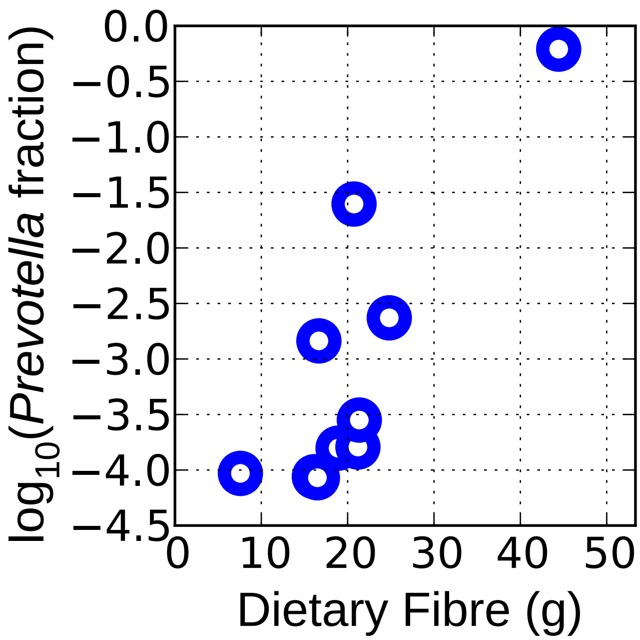
Extended Data Figure 1 | Study design. **a**, **b**, The plant-based (**a**) and animal-based (**b**) diets were fed to subjects for five consecutive days. All dates are defined relative to the start of these diet arms (day 0). Study volunteers were observed for 4 days before each diet (the baseline period, days −4 to −1) and for 6 days after each diet arm (the washout period, days 5 to 10) in order to measure subjects' eating habits and assess their recovery from each diet arm. Subjects were instructed to eat normally during both the baseline and washout periods. Stool samples were collected daily on both diet arms and 16S rRNA and fungal ITS sequencing was performed on all available samples. Subjects also kept daily diet logs. Several analyses (RNA-seq, SCFAs and bile acids) were performed primarily using only two samples per person per diet (that is, a baseline and diet arm comparison). Comparative sampling did not always occur using exactly the same study days owing to limited sample availability for some subjects. Because we expected the animal-based diet to promote ketogenesis, we only measured urinary ketones on the animal-based diet. To test the hypothesis that microbes from fermented foods on the animal-based diet survived transit through the gastrointestinal tract, we cultured bacteria and fungi before and after the animal-based diet.



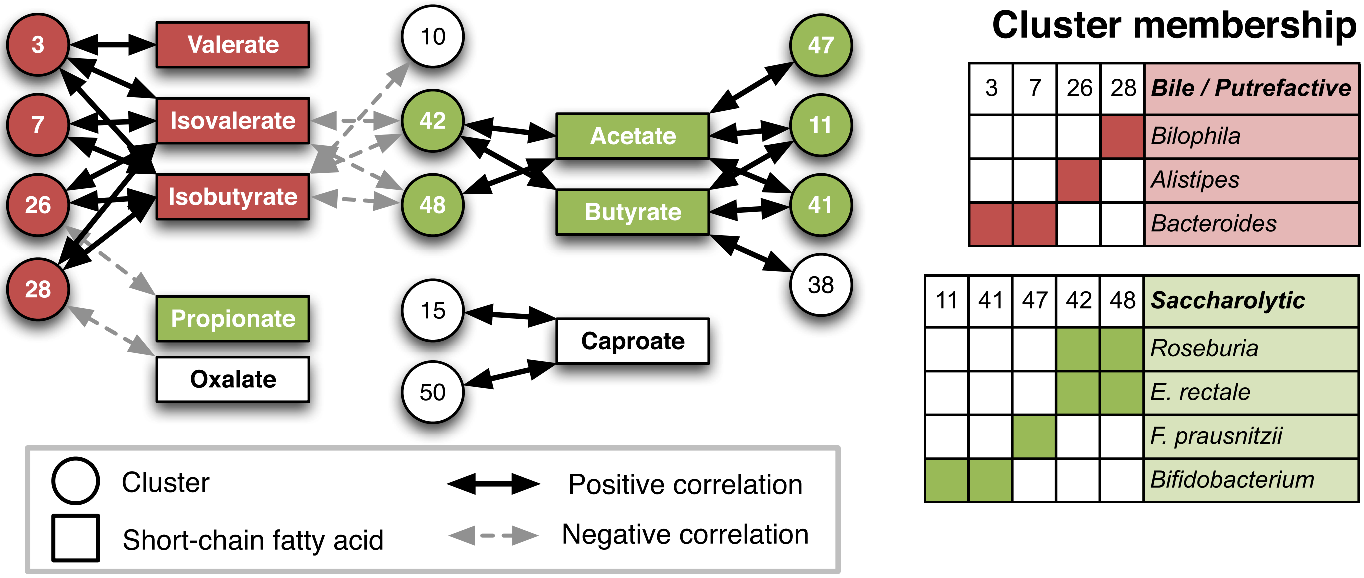
Extended Data Figure 2 | A vegetarian's microbiota. **a**–**c**, One of the study subjects is a lifelong vegetarian (subject 6). **a**, Relative abundances of *Prevotella* and *Bacteroides* are shown across the plant-based diet for subject 6 (orange circles), as well as for all other subjects (green circles). Consecutive daily samples from subject 6 are linked by dashed lines. For reference, median baseline abundances are depicted using larger circles. **b**, Relative abundances are also shown for samples taken on the animal-based diet. Labelled points correspond to diet days where subject 6's gut microbiota exhibited an increase in the relative abundance of *Bacteroides*. **c**, A principal-coordinates-based characterization of overall community structure for subject 6, as well as all other subjects. QIIME[30] was used to compute microbial β diversity with the Bray–Curtis, unweighted UniFrac and weighted UniFrac statistics. Sample similarities were projected onto two dimensions using principal coordinates analysis. Top, when coloured by subject, samples from subject 6 (green triangles) partition apart from the other subjects' samples. Bottom, of all of subject 6's diet samples, the ones most similar to the other subjects' are the samples taken while consuming the animal-based diet.



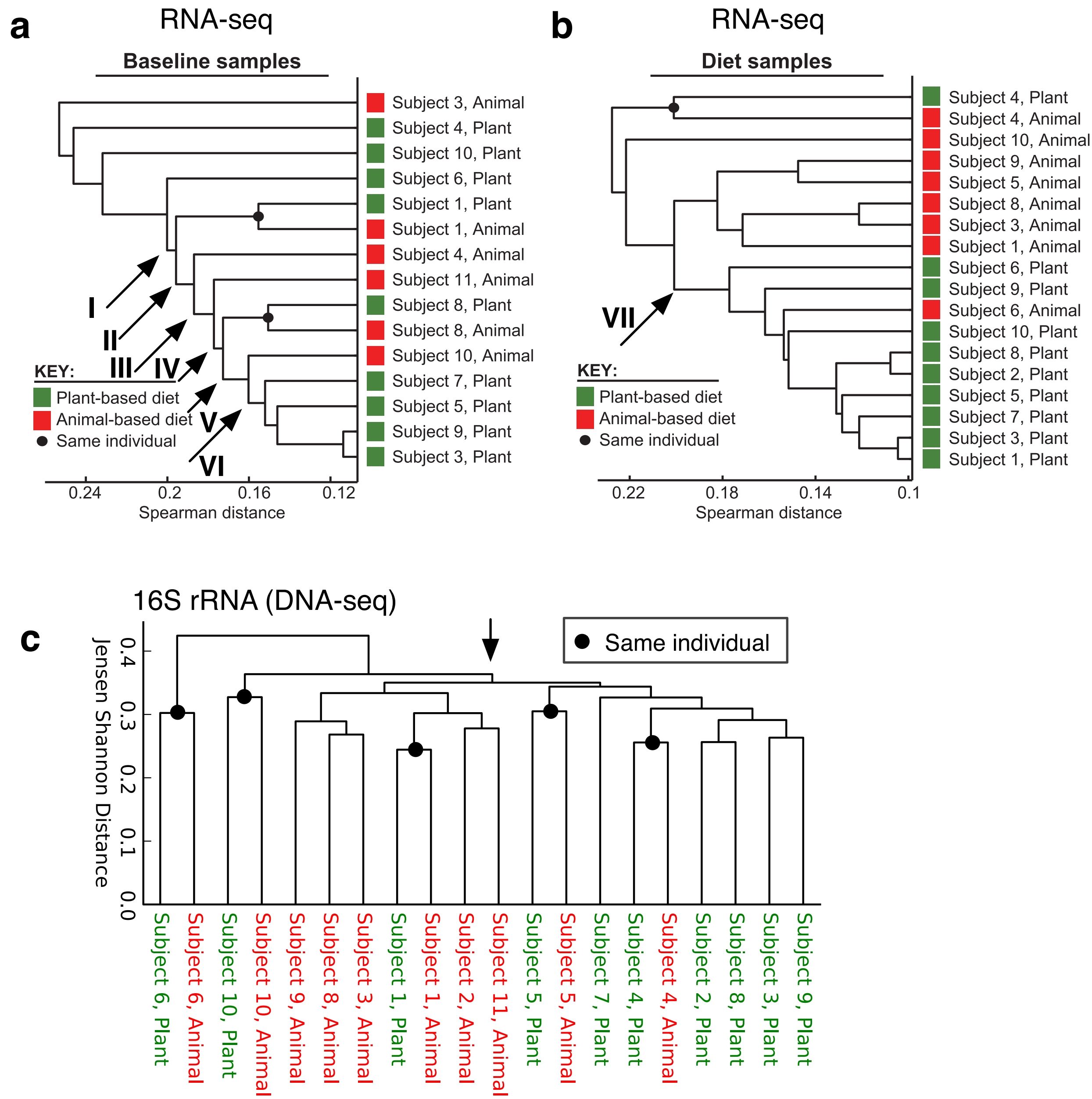
Extended Data Figure 3 | Subject physiology across diet arms.
a, Gastrointestinal motility, as measured by the initial appearance of a non-absorbable dye added to the first and last lunch of each diet. The median time until dye appearance is indicated with red arrows. Subject motility was significantly lower ($P < 0.05$, Mann–Whitney U test) on the animal-based diet (median transit time of 1.5 days) than on the plant-based one (1.0 days). **b**, Range (shaded boxes) and median (solid line) of subjects' weights over time. Subjects' weight did not change significantly on the plant-based diet relative to baseline periods, but did decrease significantly on the animal-based diet (asterisks denote $q < 0.05$, Bonferroni-corrected Mann–Whitney U test). Subjects lost a median of 1.6% and 2.5% of body weight by days 3 and 4, respectively, of the animal-based diet arm. **c**, Measurements of subjects' urinary ketone levels. Individual subjects are shown with black dots, and median values are connected with a black solid line. Urinary ketone readings were taken from day 0 of the animal-based diet onwards. Ketone levels were compared to the readings on day 0, and asterisks denote days with significant ketone increases ($q < 0.05$, Bonferroni-corrected Mann–Whitney U test; significance tests were not carried out for days on which less than four subjects reported their readings.). All subjects on the animal-based diet showed elevated levels of ketones in their urine by day 2 of the diet (≥ 15 mg dl^{-1} as compared to 0 mg dl^{-1} during initial readings), indicating that they experienced ketonuria during the diet arm. This metabolic state is characterized by the restricted availability of glucose and the compensatory extraction of energy from fat tissue[56].



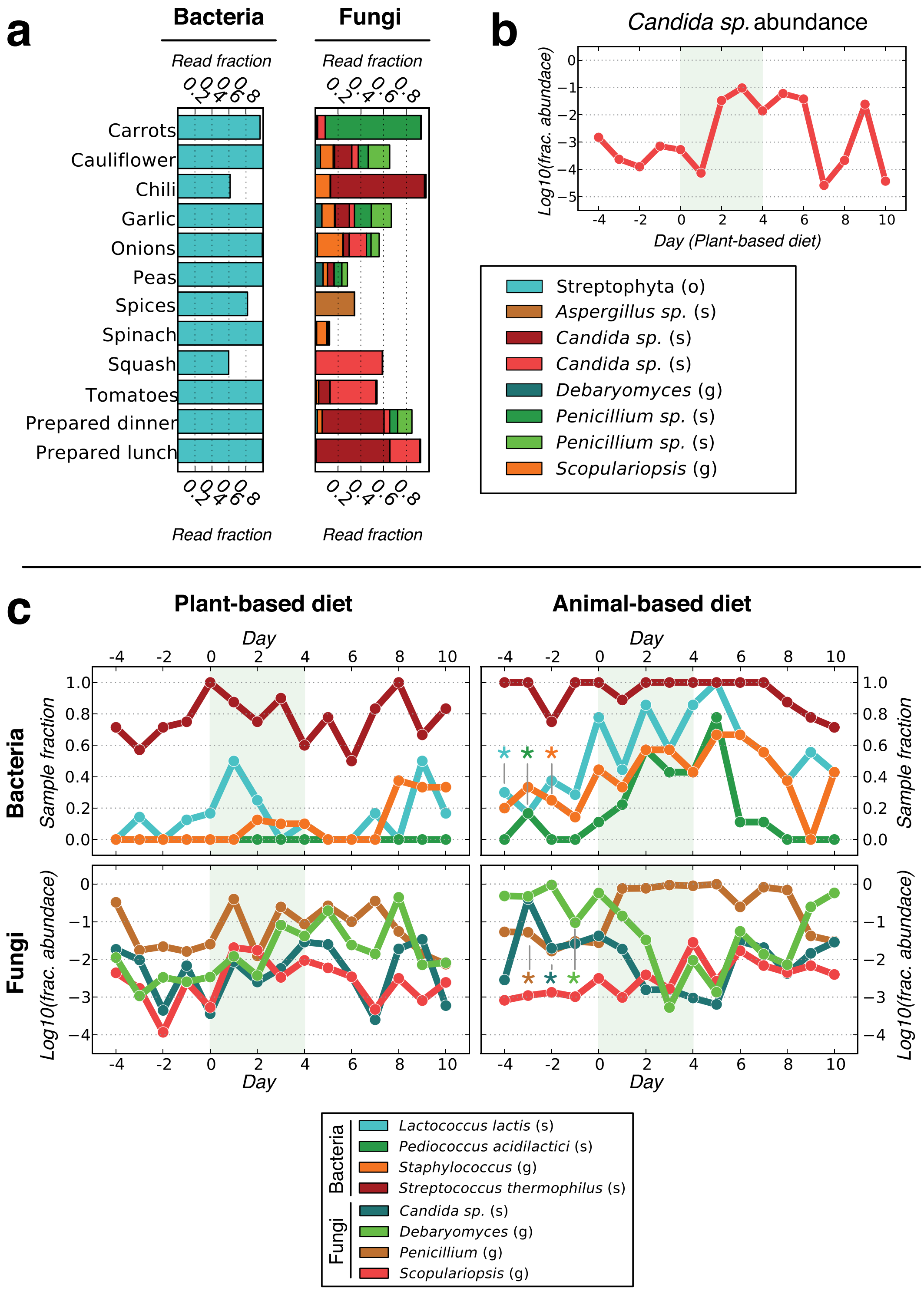
Extended Data Figure 4 | Baseline *Prevotella* abundance is associated with long-term fibre intake. *Prevotella* fractions were computed by summing the fractional 16S rRNA abundance of all OTUs whose genus name was *Prevotella*. Daily intake of dietary fibre over the previous year was estimated using the Diet History Questionnaire[32] (variable name "TOTAL_DIETARY_FIBER_G_NDSR"). There is a significant positive correlation between subjects' baseline *Prevotella* abundance and their long-term dietary fibre intake (Spearman's $\rho = 0.78$, $P = 0.008$).



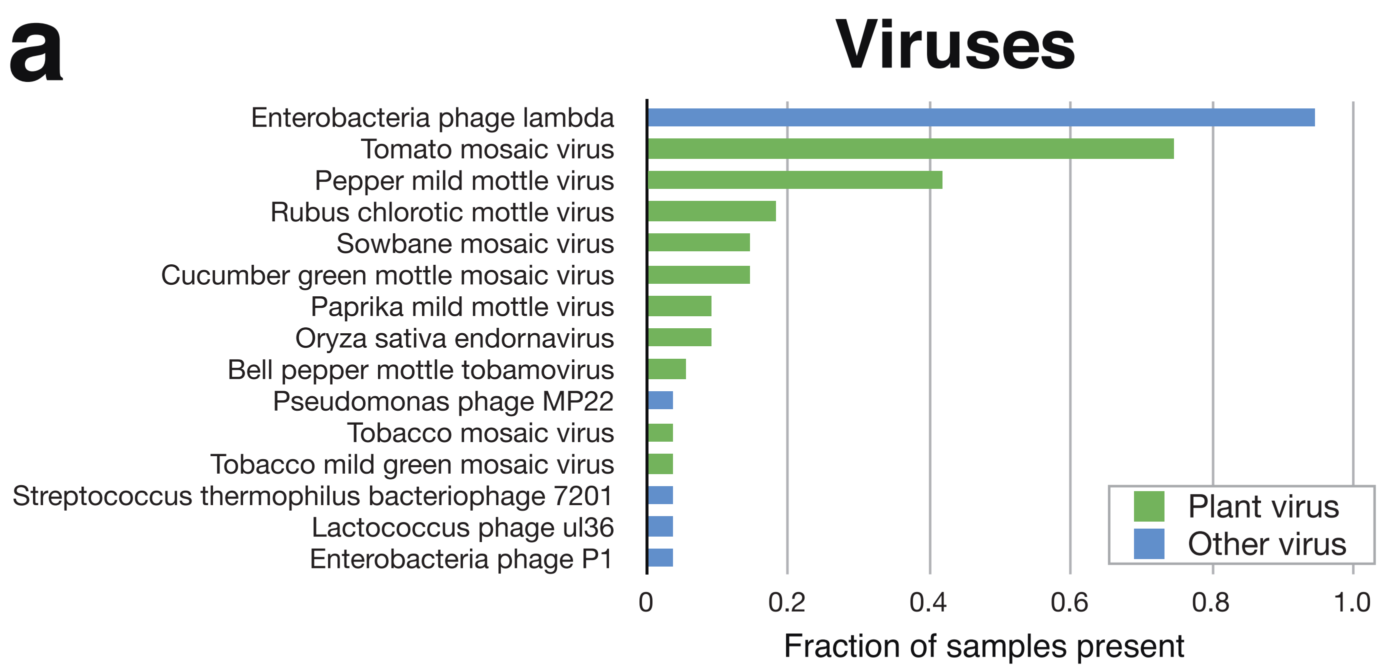
Extended Data Figure 5 | Significant correlations between SCFAs and cluster abundances across subjects. SCFAs are drawn in rectangles and coloured maroon or green if they are produced from amino acid or carbohydrate fermentation, respectively. Clusters whose members include known bile-tolerant or amino-acid-fermenting bacteria[15,16] are coloured maroon, whereas clusters including known saccharolytic bacteria[3] are coloured green. Uncoloured clusters and SCFAs are not associated with saccharolytic or putrefactive pathways. Significant positive and negative correlations are shown with black arrows and grey arrows, respectively ($q < 0.05$; Spearman correlation).

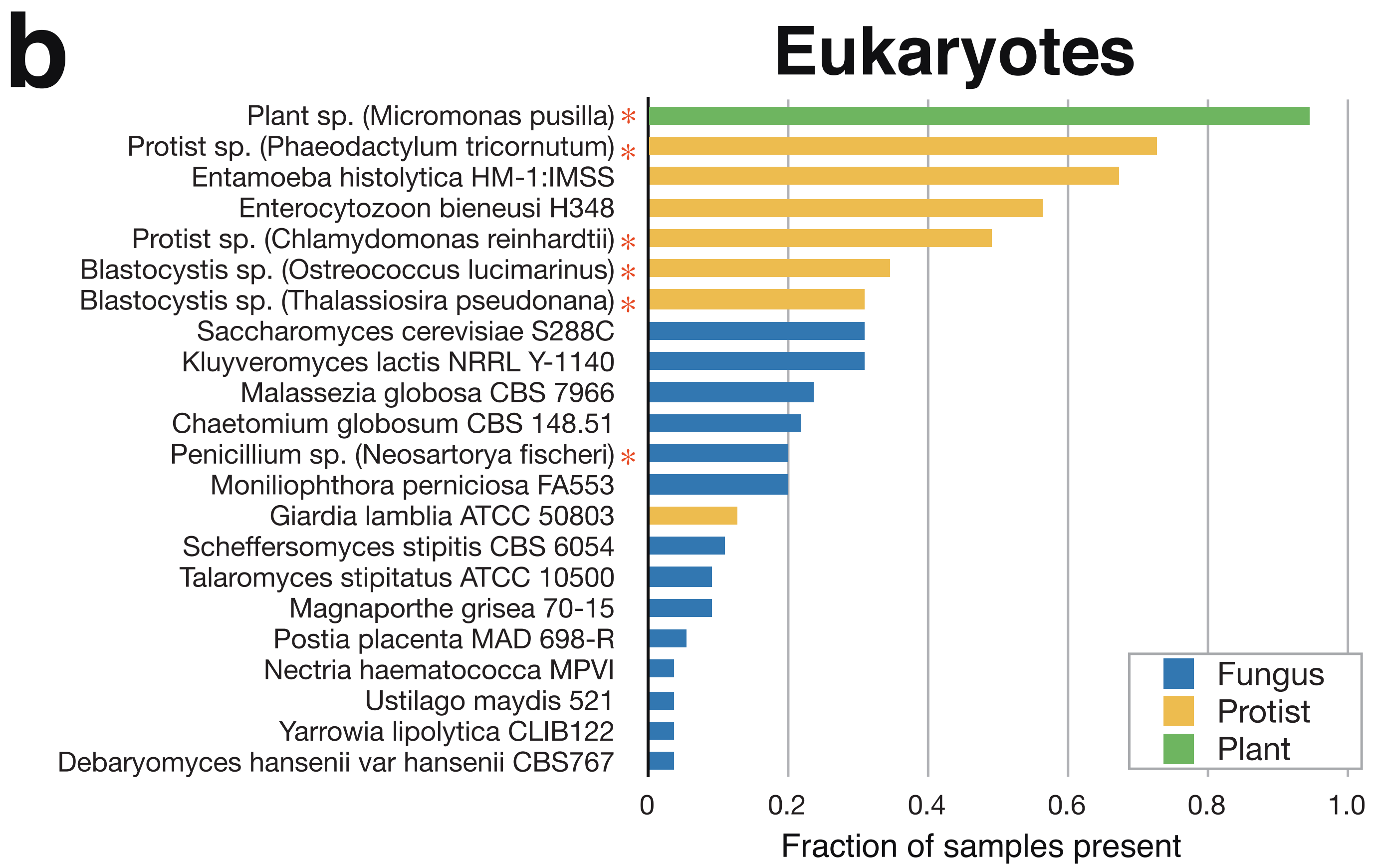


Extended Data Figure 6 | Inter-individual microbial community variation according to diet and sequencing technique. **a**, **b**, To measure the degree to which diet influences inter-individual differences in gut microbial gene expression, we clustered RNA-seq profiles from baseline (**a**) and diet (**b**) periods. Dots indicate pairs of samples that cluster by subject. The potential for diet to partition samples was measured by splitting trees at the arrowed branches and testing the significance of the resulting 2 × 2 contingency table (diet versus partition; Fisher's exact test). To avoid skewed significance values caused by non-independent samples, we only clustered a single sample per subject, per diet period. In the case of multiple baseline samples, the sample closest to the diet intervention was used. In the case of multiple diet samples, the last sample during the diet intervention was kept. A single sample was randomly chosen if there were multiple samples from the same person on the same day. No association between diet and partitioning was found for partitions I–VI ($P > 0.05$). However, a significant association was observed for partition VII ($P = 0.003$). **c**, To determine whether diet affects inter-individual differences in gut microbial community structure, we hierarchically clustered 16S rRNA data from the last day of each diet arm. Samples grouped weakly by diet: sub-trees partitioned at the arrowed node showed a minor enrichment for plant-based diet samples in one sub-tree and animal-based diet samples in the other ($P = 0.07$; Fisher's exact test). Still, samples from five subjects grouped by individual, not diet (indicated by black nodes), indicating that diet does not reproducibly overcome inter-individual differences in gut microbial community structure.

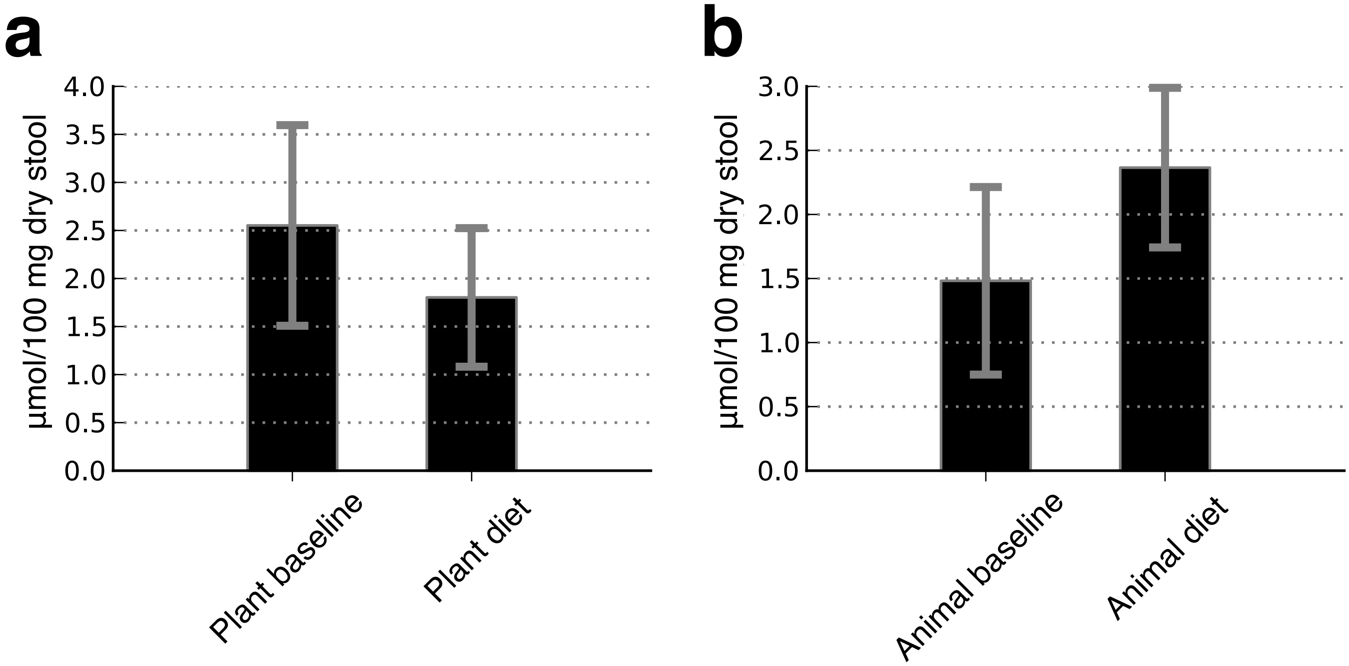


Extended Data Figure 7 | Food-associated microbes and their enteric abundance over time. **a**, Major bacterial and fungal taxa found in plant-based diet menu items were determined using 16S rRNA and ITS sequencing, respectively, at the species (s), genus (g) and order level (o). The majority of 16S rRNA gene sequences are Streptophyta, representing chloroplasts from the ingested plant matter. **b**, One of the fungi from **a**, *Candida* sp., showed a significance increase in faecal abundance on the plant-based diet ($P < 0.05$, Wilcoxon signed-rank test). **c**, Levels of bacteria and fungi associated with the animal-based diet are plotted over the plant- and animal-based diet arms. Taxa are identified on the genus (g) and species (s) level. The abundance of foodborne bacteria was near our detection limit by 16S rRNA gene sequencing; to minimize resulting measurement errors, we have plotted the fraction of samples in which bacteria are present or absent. *Lactococcus lactis*, *Pediococcus acidilactici* and *Staphylococcus*-associated reads all show significantly increased abundance on the animal-based diet ($P < 0.05$, Wilcoxon signed-rank test). Fungal concentrations were measured using ITS sequencing and are plotted in terms of log-fractional abundance. Significant increases in *Penicillium*-related fungi were observed, along with significant decreases in the concentration of *Debaryomyces* and a *Candida* sp. ($P < 0.05$, Wilcoxon signed-rank test). One possible explanation for the surprising decrease in the concentration of food-associated fungi is that the more than tenfold increase in *Penicillium* levels lowered the relative abundance of all other fungi, even those that increased in terms of absolute abundance.

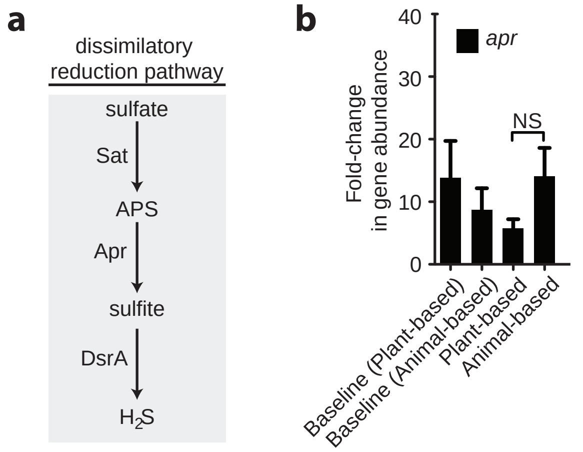




Extended Data Figure 8 | Eukaryotic and viral taxa detected via RNA-seq. **a**, Identified plant and other viruses. The most common virus is a DNA virus (lambda phage) and may be an artefact of the sequencing process. **b**, Identified fungi, protists and other eukaryotes. Taxa that were re-annotated using manually curated BLAST searches are indicated with asterisks and their original taxonomic assignments are shown in parentheses (see Methods for more details).



Extended Data Figure 9 | Faecal bile acid concentrations on baseline, plant- and animal-based diets. **a**, **b**, Median bulk bile acid concentrations are shown for all individuals on the plant-based (**a**) and animal-based (**b**) diets (error bars denote median absolute deviations). For detailed experimental protocols, see Methods. Bile acid levels did not significantly change on the plant-based diet relative to baseline levels ($P > 0.1$, Mann–Whitney U test). However, bile acid levels trended upwards on the animal-based diet, rising from 1.48 µmol per 100 mg dry stool during the baseline period to 2.37 µmol per 100 mg dry stool ($P < 0.10$, Mann–Whitney U test).



Extended Data Figure 10 | The dissimilatory sulphate reduction pathway. **a**, Microbes reduce sulphate to hydrogen sulphide by first converting sulphate to adenosine 5′-phosphosulphate (APS) via the enzyme ATP sulphurylase (Sat). Next, APS is reduced to sulphite by the enzyme APS reductase (Apr). Finally, the end product hydrogen sulphide is reached by reducing sulphite through the enzyme sulphite reductase (DsrA). This last step of the pathway can be performed by *Bilophila* and is thought to contribute to intestinal inflammation[6]. **b**, No significant changes in *apr* gene abundance were observed on any diet ($P > 0.05$, Mann–Whitney U test; $n = 10$ samples per diet arm). Values are mean ± s.e.m. However, *dsrA* abundance increased on the animal-based diet (Fig. 5d). NS, not significant.

doi:10.1038/nature12819

Glutamine methylation in histone H2A is an RNA-polymerase-I-dedicated modification

Peter Tessarz[1,2], Helena Santos-Rosa[1,2], Sam C. Robson[1,2], Kathrine B. Sylvestersen[3], Christopher J. Nelson[1,2]†, Michael L. Nielsen[3] & Tony Kouzarides[1,2]

Nucleosomes are decorated with numerous post-translational modifications capable of influencing many DNA processes[1]. Here we describe a new class of histone modification, methylation of glutamine, occurring on yeast histone H2A at position 105 (Q105) and human H2A at Q104. We identify Nop1 as the methyltransferase in yeast and demonstrate that fibrillarin is the orthologue enzyme in human cells. Glutamine methylation of H2A is restricted to the nucleolus. Global analysis in yeast, using an H2AQ105me-specific antibody, shows that this modification is exclusively enriched over the 35S ribosomal DNA transcriptional unit. We show that the Q105 residue is part of the binding site for the histone chaperone FACT (facilitator of chromatin transcription) complex[2]. Methylation of Q105 or its substitution to alanine disrupts binding to FACT *in vitro*. A yeast strain mutated at Q105 shows reduced histone incorporation and increased transcription at the ribosomal DNA locus. These features are phenocopied by mutations in FACT complex components. Together these data identify glutamine methylation of H2A as the first histone epigenetic mark dedicated to a specific RNA polymerase and define its function as a regulator of FACT interaction with nucleosomes.

Glutamine methylation occurs on translation termination factors and ribosomal proteins[3]. We investigated whether such a modification exists on histones by interrogating mass spectrometric data sets. We identified a single glutamine, human Q104 (yeast: Q105) in H2A as a site of methylation (Fig. 1a and Extended Data Fig. 1). The residue is located on the surface of the octamer (Extended Data Fig. 2) and is highly conserved in canonical H2A from yeast to human. However, in H2A.Z it is exchanged for a glycine or serine (Fig. 1b). We raised a modification-specific antibody (Extended Data Fig. 3) that detects this modification in yeast and mammalian cells (Fig. 1c).

To identify the methyltransferase responsible, we performed a candidate approach and screened 72 predicted yeast non-essential predicted methyltransferases[4] by analysing knockout lysates by western blotting with the modification-specific antibody. However, we did not detect loss of signal in any lysate (not shown). We then used an unbiased biochemical approach and fractionated yeast cells as described in Fig. 2a. Fractions were assayed on a 20-residue peptide spanning Q105, or the respective QA mutant, coupled to beads in the presence of tritiated *S*-adenosyl-methionine (SAM). Methyltransferase activity was assessed by scintillation counting (Fig. 2b) and the fraction containing activity towards H2AQ105 was subjected to mass spectrometry (Supplementary Table 4). All 178 non-essential proteins identified by mass spectrometry were tested by knockout analysis and western blotting, but none showed reduction of Q105 methylation (not shown). We then examined the essential proteins in the active fraction. We focused on Nop1, a known rRNA methyltransferase[5], because its essential co-factors, Nop56/58 (ref. 6) and other members of an active RNA polymerase I complex[7] were present in this active fraction. Furthermore, with the exception of Nop58, all the previously mentioned proteins are known to interact with H2A[8,9]. Tandem affinity purification (TAP)-tagging of the Nop1 complex purifies an enzymatic activity that methylates H2AQ105 (Fig. 2c). Next, recombinantly purified Nop1 was tested for its ability to modify recombinant H2A *in vitro*. Indeed, Nop1, in the presence of H2A and SAM, methylates H2AQ105 as detected by the H2AQ105me-specific antibody (Fig. 2d and Extended Data Fig. 4a). Additionally, mass spectrometry of this reaction identifies H2AQ105 methylation (Fig. 2e and Extended

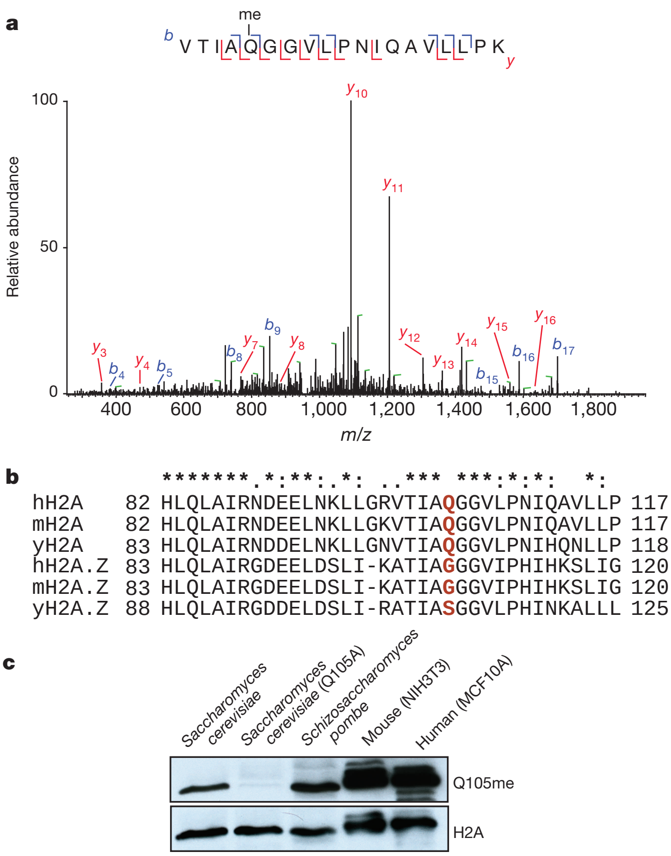


Figure 1 | Identification and localization of methylated H2A glutamine 105. a, Tandem mass spectrum of the Q104me modified peptide VTIAQGGVLPNIQAVLLPK from H2A. The *y* and *b* series indicate fragments at amide bonds of the peptide, unambiguously identifying the methylated glutamine 104 (in mammalian cells; in yeast, glutamine 105). **b**, Alignment of the region encompassing Q105 of H2A and its variant H2A.Z. Highlighted in red is Q105 in H2A and the corresponding change to glycine or serine in H2A.Z. **c**, Analysis of yeast and mammalian cell extracts for the presence of Q105 methylation using a modification-specific antibody.

[1]Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK. [2]Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK. [3]Department of Proteomics, The Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark. †Present address: Department of Biochemistry and Microbiology, University of Victoria, 3800 Finnerty Road, Victoria, British Columbia V8P 5C2, Canada.

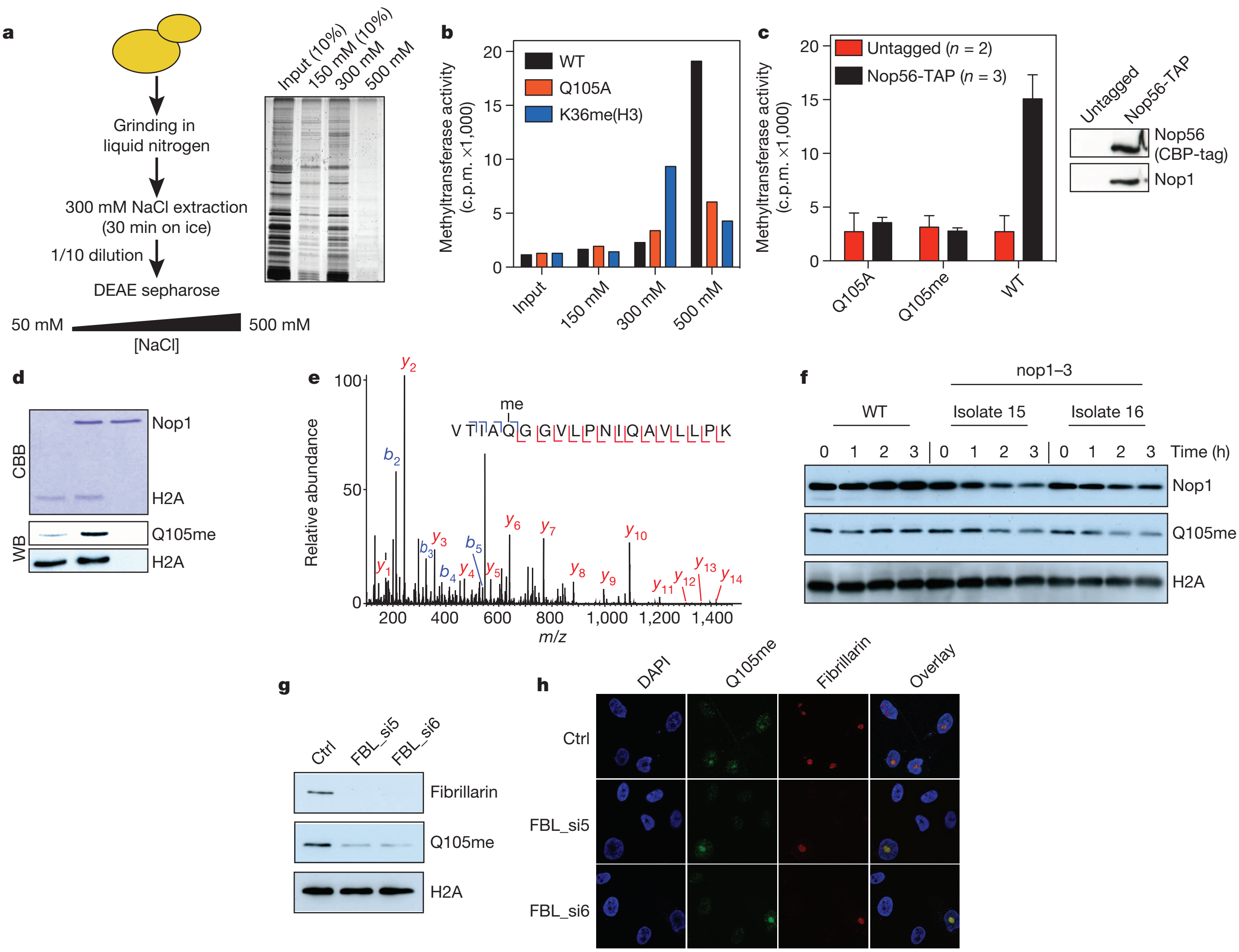


Figure 2 | Identification of Nop1/fibrillarin as the methyltransferase of Q105. **a**, General strategy. **b**, Fractions from **a** were assayed on peptides containing glutamine or alanine at position 105 and compared with an unrelated peptide (H3K36me). For this, peptides were bound to Dynabeads, incubated with extract and tritiated SAM and, after extensive washes, analysed by liquid scintillation. Representative data of three independent experiments are shown. **c**, TAP-tag purification of the Nop1 complex coupled to the same activity assay as in **b** recapitulates the activity as found in the DEAE fraction. **d**, Recombinant Nop1 was incubated with SAM and recombinant histone H2A. Coomassie stain (CBB) and western blot (WB) of the reaction are shown. **e**, Tandem mass spectrum of the Q104me modified peptide from H2A, which unambiguously identifies the methylated glutamine 104. **f**, Strains carrying thermosensitive alleles (15 and 16) of Nop1 were analysed for loss of Q105 methylation levels upon shift to restrictive temperature. **g**, The mammalian homologue of Nop1, fibrillarin, was knocked down by independent siRNAs and probed for loss of Q105 methylation 48 h after transfection. A scrambled siRNA served as control (Ctrl). **h**, Immunofluorescence of cells treated as in **g** were stained using the Q105me-specific and anti-fibrillarin antibodies and counterstained with 4′,6-diamidino-2-phenylindole (DAPI) as a nuclear marker.

Data Fig. 4b). To test the enzymatic activity of Nop1 *in vivo* we made use of two independently isolated thermosensitive mutants carrying the same amino-acid changes, which are located in the SAM binding site of Nop1 (ref. 5). Yeast harbouring these thermosensitive (*ts*) alleles showed a 50% reduced Q105 methylation signal upon shift to restrictive temperature at a time at which cells are still proliferating (Fig. 2f and Extended Data Fig. 5a, b). These results identify Nop1 as the enzyme responsible for H2AQ105 methylation in yeast.

Nop1 has a single highly conserved homologue in human cells, called fibrillarin[10] (Extended Data Fig. 5c, d). To establish that fibrillarin methylates Q104 in human cells, it was knocked down in MCF10A cells. Transfection of two independent short interfering RNAs (siRNAs) against fibrillarin leads to robustly reduced amounts of H2AQ104me (Fig. 2g and Extended Data Fig. 5e). At this time viability was only marginally affected, based on MTT proliferation assays (Extended Data Fig. 5f). Furthermore, immunofluorescence showed that Q104me and fibrillarin were enriched in the nucleolus of MCF10A cells (Fig. 2h). However, siRNA-induced knockdown of fibrillarin completely abrogated detection of Q104me in the nucleolus. We observed no morphological changes of the nucleus and nucleolus that have been reported to occur upon prolonged fibrillarin knockdown[11], indicating that—in agreement with the MTT assay—the viability of the cells was not affected at the time of analysis.

The main function of the nucleolus is ribosomal DNA (rDNA) transcription and ribosome biogenesis[12]. To analyse the distribution of H2A glutamine methylation in chromatin, we performed chromatin immunoprecipitations coupled to deep DNA sequencing (ChIP-seq). The rDNA locus consists of roughly 100–200 repeats in yeast and 200–400 copies in human cells[13], of which about half are active and almost devoid of nucleosomal structure and the other half are inactive and densely packed with nucleosomes[13,14]. In yeast up to about 80% of rDNA repeats can be deleted, in which case all the remaining repeats, approximately 20 copies, are active[15]. This strain still retains H2AQ105me (Extended Data Fig. 6). Remarkably, when the H2AQ105me antibody

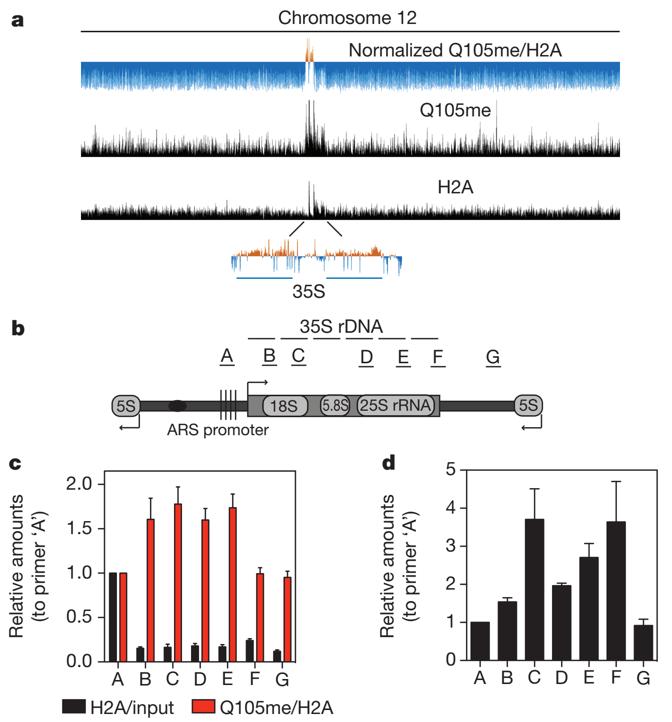


Figure 3 | Genomic landscape of Q105me. **a**, ChIP-seq profile of Q105me and H2A over chromosome 12. The upper plot represents the normalized tracks for the mapped reads of Q105me/H2A. The magnification of the enriched region shows the rDNA region located on this chromosome. The two blue bars represent the 35S transcripts that are present in the genome annotation. **b**, Representation of one rDNA repeat with the position of primers used to scan the rDNA region by ChIP–quantitative (q)PCR. **c**, ChIP–qPCR validation of the ChIP-seq shown in **a**. **d**, Nop1 profile over the rDNA locus. The ChIP–qPCR profile was internally normalized to signal of primer pair 'A'. ChIP–qPCR data show the mean ± s.e.m. of three independent biological experiments.

is used in ChIP-seq analysis, the only site of enrichment in the entire yeast genome is over the 35S rDNA transcription units (Fig. 3a, c). In addition, Nop1, the enzyme that mediates H2AQ105 methylation, co-localizes with H2AQ105 methylation at the 35S rDNA locus (Fig. 3d). Together, these results indicate that both in human and in yeast, methylation of H2A represents a modification restricted to the nucleolus. Given the enrichment of glutamine methylation on the transcribed region of the rDNA cluster, we asked whether RNA polymerase I transcription was required for deposition of Q105 methylation. We used actinomycin D at concentrations known to inhibit RNA polymerase I, but not RNA polymerase II[16]. These conditions led to reduced Q104/5 methylation in mammalian/yeast cells, indicating that active RNA Pol I transcription is required for glutamine methylation to occur (Extended Data Fig. 7).

Histone methylation can act as a platform to recruit and regulate other chromatin-related factors[1] such as chromatin remodelling complexes. The region spanning Q105 in H2A has previously been described as a potential binding site for FACT[17,18], a protein complex consisting of Spt16/Pob3 and Nhp6a in yeast, which is required for efficient passage of RNA and DNA polymerases through chromatin by remodelling nucleosomes[2]. Residues in the region of H2A spanning Q105 show genetic interactions with FACT *ts* mutants[17,18], as does a Q105A mutant using a transcription-based reporter system (Extended Data Fig. 8).

To probe the possibility that FACT physically interacts with H2A through the region spanning Q105, we took an unbiased approach (phage display) to identify regions in H2A/H2B interacting with Spt16 and Pob3. A randomized 12-residue peptide phage library was used to enrich for sequences binding to Spt16/Pob3. In line with the published genetic findings, the interaction screen identified a consensus sequence spanning H2AQ105 as the binding site for FACT (Extended Data Fig. 9a).

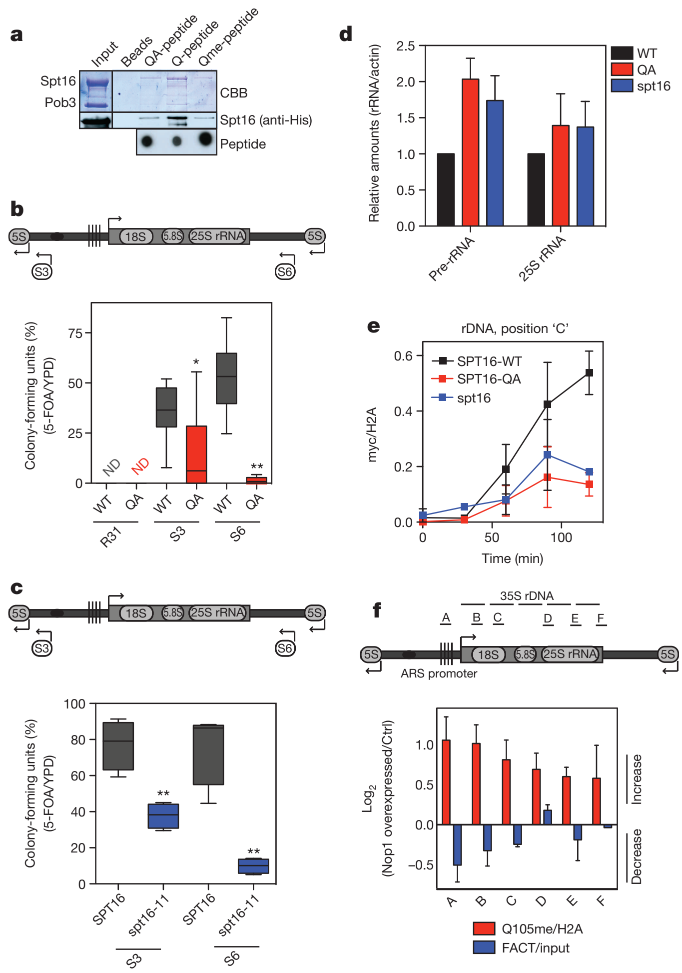


Figure 4 | Unmodified Q105 is part of a recognition motif for FACT. **a**, Indicated peptides were bound to streptavidin-coupled Dynabeads and incubated with recombinantly purified Spt16/Pob3 for 4 h at 4 °C. Input represents 50% of Spt16/Pob3 used for the immunoprecipitations. Bound Spt16/Pob3 was analysed by Coommassie stain (CBB) or western blotting. Peptides were spotted on membrane as loading control. **b**, Effect of the H2AQ105A mutant on the transcription of a *URA3* reporter integrated at the indicated positions in the rDNA repeat (S3 and S6) and R31 outside the rDNA[19]. Strains were assayed for their ability to grow on 5-FOA for 3 days at 30 °C ($n \geq 3$; ND, not detected). **c**, The *spt16-11* allele was assayed as described in **c**, but colonies were counted after 1 week. **d**, Yeast expressing either wild-type (WT) or Q105A (QA) mutant H2A as the sole source of histones were assayed for rDNA transcription by transcriptional run-on and compared with cells carrying an *spt16* thermosensitive allele (*spt16*). Amounts of RNA are expressed relative to actin. **e**, Differences in histone incorporation rates between WT and Q105A (QA) histones. WT H2A and H2AQ105 were myc-tagged and placed under the control of a galactose-inducible promoter in either wild-type or *spt16-11* cells. Myc-tagged histones were induced by addition of 2% galactose ($t = 0$). Samples were taken at the indicated time points and chromatin immunoprecipitated with anti-myc and H2A antibodies. **f**, Nop1 overexpression (Nop1 overexpressed) leads to an increase of Q105 methylation and a concurrent decrease in FACT occupancy at the rDNA locus. Primer positions are indicated and data presented as the $\log_2$ changes compared with an empty vector control (Ctrl). ChIP–qPCR data show the mean ± s.e.m. of at least two independent biological experiments. $^*P<0.05$; $^{**}P<0.01$.

We next asked whether methylation of H2AQ105 could influence the binding to FACT. Figure 4a shows that binding of recombinant Spt16/Pob3 to a peptide spanning H2AQ105 is significantly decreased when Q105 is methylated or mutated to alanine. Pull-downs from HeLa nuclear extracts using the same peptides demonstrate that the endogenous human FACT complex is responsive to glutamine methylation on H2A (Extended Data Fig. 9b), suggesting that the disruption

of FACT binding to this site is the mechanistic consequence of glutamine methylation.

We next sought to explore the consequence of H2AQ105 methylation on FACT function *in vivo*. To do this, we took advantage of the fact that mutation of H2AQ105 to alanine phenocopies Q105 methylation in terms of effecting FACT binding (Fig. 4a). To study a potential influence of Q105 methylation on rDNA transcription and its interplay with FACT, we turned to a well-established reporter-based system that allows sensitive monitoring of the transcriptional state of the rDNA locus, in which weak, but constitutively expressed, *URA3* cassettes were integrated into the rDNA locus (S3 and S6; Fig. 4b)[19]. Figure 4b shows a significant drop in 5-FOA-positive colonies in the Q105A mutant, indicative of a higher transcription rate of the *URA3*. We then introduced the *spt16-11* allele—which leads to a 30–40% decrease in FACT protein levels[17]—into the same reporter strains. Figure 4d shows that at semi-permissive temperature we observe a drastic loss of colony numbers on 5-FOA plates, pointing to an increase in transcriptional permissiveness (open chromatin) in accordance with recently published findings[20]. These results show that reduced FACT activity indeed phenocopies the Q105A mutation. Thus, a mutation that disrupts the function of the chromatin remodeller FACT or a mutation that disrupts the binding of FACT to chromatin lead to transcriptional permissiveness at the rDNA locus. To monitor RNA Pol I transcription rates directly, we performed run-on experiments in wild-type, FACT *ts* and Q105A strains. Figure 4d shows that rDNA transcription was increased using two different primer pairs, confirming a more permissive chromatin at the rDNA locus, when FACT or Q105A are mutated.

One possible explanation for the increased rDNA transcription in the Q105A and FACT *ts* strains, is loss of nucleosomes over a transcribed region. FACT *ts* mutants have already been described as possessing such a phenotype[21]. To investigate a possible histone deposition defect, we generated strains in which we placed myc-tagged versions of either wild-type H2A or the Q105A mutant under control of the *GAL1*-promoter in a wild-type or *spt16-11* yeast background. Induction of the myc-tagged histones was identical as judged by total steady-state levels on western blots (Extended Data Fig. 10a). The wild-type H2A was very efficiently deposited into chromatin as monitored by ChIP (Fig. 4e). However, the *spt16-11* and the Q105A mutant had a profound defect in H2A incorporation into chromatin (Fig. 4e). These findings suggest that methylation of H2AQ105, as phenocopied by the H2AQ105A mutation and a FACT *ts* mutant, results in the transcriptional stimulation of the *URA3* reporters because of reduced nucleosomal occupancy in the rDNA repeat.

Finally, we set out to test directly whether an increase in Q105 methylation on the rDNA locus would decrease FACT occupancy. Indeed, overexpression of Nop1 leads to increased Q105me and is accompanied by a decrease of FACT occupancy, in line with the hypothesis that Q105 methylation is regulating FACT availability on chromatin (Fig. 4f).

The findings presented here identify a new histone modification pathway operational exclusively in the nucleolus. It involves methylation of H2AQ105 by Nop1 in yeast, and methylation of H2AQ104 by fibrillarin in human cells, resulting in the weakening of interactions between H2A and FACT. The FACT complex interacts with all three RNA polymerases[22,23] and facilitates transcription in two steps by (1) binding and disrupting nucleosomes in the path of the polymerase[22,24,25] and (2) by augmenting the re-deposition of nucleosomes in the wake of transcribing polymerase[24,26]. Glutamine methylation of H2A may affect either of these functions by disrupting binding to FACT. The observation that an H2AQ105A mutant is incorporated to a lower extent would favour a model in which re-deposition is decreased. Such a model is also in agreement with earlier observations that rDNA has a low nucleosome occupancy[13], in contrast to most other regions of the genome. Indeed, recent reports suggest that H2A in particular seems to be depleted from this region[27] and that FACT might play a role in this pathway[20]. The net result of a glutamine-modified chromatin state is that RNA Pol I transits less impeded by nucleosomes. It is worth noting, however, that glutamine methylation of H2A is present in the rDNA locus, even though deposition is affected. The residual loading of modified H2A might be due to the presence of other histone chaperones such as Nap1 that are insensitive to glutamine methylation on H2A (Extended Data Fig. 10b).

RNA Pol I associates with the Nop1 enzyme and may carry it along during transcription elongation. It is currently unclear how glutamine methylation is initiated or reversed, but it might be linked to the reactivation of the rDNA locus that has been described to occur after DNA replication in an RNA Pol-I-dependent manner[27]. Another possibility is that the enzyme itself is the key node of regulation. Nop1/fibrillarin is highly modified[28,29], so a signalling pathway leading to the glutamine methyltransferase could be the triggering event.

Glutamine methylation of H2A represents the first histone modification that is dedicated to only one of the three RNA polymerases. The selectivity for Pol I and its compartmentalization within the nucleolus might be necessary to generate a chromatin state capable of dealing with the high demands for transcription of ribosomal components. Indeed, glutamine methylation as a whole seems to be a modification that is dedicated to ribosomal biosynthesis: Nop1 has the ability also to methylate rRNA and affect RNA processing[5]; the only other known glutamine methyltransferases in yeast (Mtq1 and Mtq2) modify translational release factors on a conserved glutamine[3]. Thus, glutamine methylation may have evolved to be a modification dedicated to a specific cellular process. Finally, our finding that a protein can catalyse the methylation of proteins and RNA opens the possibility that many other enzymes may have such dual specificity.

METHODS SUMMARY

The antibody against H2AQ105me was raised using the speedy 28-day programme of Eurogentec using modified peptides coupled to KLH. Yeast genetics, molecular biology, cell culture and biochemistry were performed using standard methods. ChIP was essentially as described earlier[30]. ChIP-seq was analysed on an Illumina MiSeq. Mapped ChIP-seq reads were normalized by dividing Q105me counts by the H2A counts. Mass spectrometry was performed upon in-gel digest on a LTQ–Orbitrap XL (Thermo Fisher Scientific) and analysed using the MaxQuant software package. Detailed information about the reagents and methodology used is available in Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 January; accepted 29 October 2013.
Published online 18 December 2013.

1. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
2. Formosa, T. The role of FACT in making and breaking nucleosomes. *Biochim. Biophys. Acta* **1819,** 247–255 (2012).
3. Polevoda, B. & Sherman, F. Methylation of proteins involved in translation. *Mol. Microbiol.* **65,** 590–606 (2007).
4. Petrossian, T. C. & Clarke, S. G. Multiple motif scanning to identify methyltransferases from the yeast proteome. *Mol. Cell. Proteomics* **8,** 1516–1526 (2009).
5. Tollervey, D., Lehtonen, H., Jansen, R., Kern, H. & Hurt, E. C. Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly. *Cell* **72,** 443–457 (1993).
6. Gautier, T., Bergès, T., Tollervey, D. & Hurt, E. Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis. *Mol. Cell. Biol.* **17,** 7088–7098 (1997).
7. Fath, S. *et al.* Association of yeast RNA polymerase I with a nucleolar substructure active in rRNA synthesis and processing. *J. Cell Biol.* **149,** 575–590 (2000).
8. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae. Nature* **440,** 637–643 (2006).
9. Lambert, J.-P., Mitchell, L., Rudner, A., Baetz, K. & Figeys, D. A novel proteomics approach for the discovery of chromatin-associated protein networks. *Mol. Cell. Proteomics* **8,** 870–882 (2009).
10. Jansen, R. P. *et al.* Evolutionary conservation of the human nucleolar protein fibrillarin and its functional expression in yeast. *J. Cell Biol.* **113,** 715–729 (1991).
11. Amin, M. A. *et al.* Fibrillarin, a nucleolar protein, is required for normal nuclear morphology and cellular growth in HeLa cells. *Biochem. Biophys. Res. Commun.* **360,** 320–326 (2007).
12. Raška, I., Shaw, P. J. & Cmarko, D. Structure and function of the nucleolus in the spotlight. *Curr. Opin. Cell Biol.* **18,** 325–334 (2006).
13. Albert, B., Perez-Fernandez, J., Léger-Silvestre, I. & Gadal, O. Regulation of ribosomal RNA production by RNA polymerase I: does elongation come first? *Genet. Res. Int.* **2012,** 276948 (2012).

14. Grummt, I. & Längst, G. Epigenetic control of RNA polymerase I transcription in mammalian cells. *Biochim. Biophys. Acta* **1829,** 393–404 (2013).
15. Ide, S., Miyazaki, T., Maki, H. & Kobayashi, T. Abundance of ribosomal RNA gene copies maintains genome integrity. *Science* **327,** 693–696 (2010).
16. Gorenstein, C., Atkinson, K. D. & Falkes, E. V. Isolation and characterization of an actinomycin D-sensitive mutant of *Saccharomyces cerevisiae. J. Bacteriol.* **136,** 142–147 (1978).
17. VanDemark, A. P. *et al.* Structural and functional analysis of the Spt16p N-terminal domain reveals overlapping roles of yFACT subunits. *J. Biol. Chem.* **283,** 5058–5068 (2008).
18. McCullough, L. *et al.* Insight into the mechanism of nucleosome reorganization from histone mutants that suppress defects in the FACT histone chaperone. *Genetics* **188,** 835–846 (2011).
19. Smith, J. S. & Boeke, J. D. An unusual form of transcriptional silencing in yeast ribosomal DNA. *Genes Dev.* **11,** 241–254 (1997).
20. Johnson, J. M. *et al.* Rpd3 and Spt16-mediated nucleosome assembly and transcriptional regulation on yeast rDNA genes. *Mol. Cell. Biol.* **33,** 2748–2759 (2013).
21. Hainer, S. J., Pruneski, J. A., Mitchell, R. D., Monteverde, R. M. & Martens, J. A. Intergenic transcription causes repression by directing nucleosome assembly. *Genes Dev.* **25,** 29–40 (2011).
22. Orphanides, G., Wu, W. H., Lane, W. S., Hampsey, M. & Reinberg, D. The chromatin-specific transcription elongation factor FACT comprises human SPT16 and SSRP1 proteins. *Nature* **400,** 284–288 (1999).
23. Birch, J. L. *et al.* FACT facilitates chromatin transcription by RNA polymerases I and III. *EMBO J.* **28,** 854–865 (2009).
24. Belotserkovskaya, R. *et al.* FACT facilitates transcription-dependent nucleosome alteration. *Science* **301,** 1090–1093 (2003).
25. Winkler, D. D., Muthurajan, U. M., Hieb, A. R. & Luger, K. Histone chaperone FACT coordinates nucleosome interaction through multiple synergistic binding events. *J. Biol. Chem.* **286,** 41883–41892 (2011).
26. Formosa, T. *et al.* Defects in SPT16 or POB3 (yFACT) in *Saccharomyces cerevisiae* cause dependence on the Hir/Hpc pathway: polymerase passage may degrade chromatin structure. *Genetics* **162,** 1557–1571 (2002).
27. Wittner, M. *et al.* Establishment and maintenance of alternative chromatin states at a multicopy gene locus. *Cell* **145,** 543–554 (2011).
28. Choudhary, C. *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325,** 834–840 (2009).
29. Wang, B., Malik, R., Nigg, E. A. & Körner, R. Evaluation of the low-specificity protease elastase for large-scale phosphoproteome analysis. *Anal. Chem.* **80,** 9526–9533 (2008).
30. Santos-Rosa, H. *et al.* Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. *Mol. Cell* **12,** 1325–1332 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. Hurt, D. Stillman, T. Kobayashi, J. Smith and J. Boeke for providing strains; K. Lilley for mass spectrometry and M. Vermeulen for help with identifying H2AQ104me; S. Moss for running the Illumina MiSeq; members of the Kouzarides laboratory for discussions; and A. Bannister and R. Belotserkovskaya for reading the manuscript. This work has been financed by a programme grant from Cancer Research UK and a project grant from the Biotechnology and Biological Sciences Research Council (BB/K017438/1).

Author Contributions P.T. and H.S.-R. designed experiments, performed research, interpreted data and wrote the manuscript. K.B.S. and M.L.N. performed mass spectrometry. C.J.N. supplied new reagents. T.K. designed experiments, interpreted data and wrote the manuscript.

Author Information Data of the ChIP-seq experiments have been deposited in Array Express under accession number E-MTAB-1447. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.K. (t.kouzarides@gurdon.cam.ac.uk).

doi:10.1038/nature12811

Structural basis of the alternating-access mechanism in a bile acid transporter

Xiaoming Zhou[1,2]*, Elena J. Levin[1]*, Yaping Pan[1], Jason G. McCoy[1], Ruchika Sharma[2], Brian Kloss[3], Renato Bruni[3], Matthias Quick[4,5] & Ming Zhou[1,2,6]

Bile acids are synthesized from cholesterol in hepatocytes and secreted through the biliary tract into the small intestine, where they aid in absorption of lipids and fat-soluble vitamins. Through a process known as enterohepatic recirculation, more than 90% of secreted bile acids are then retrieved from the intestine and returned to the liver for resecretion[1]. In humans, there are two Na^+-dependent bile acid transporters involved in enterohepatic recirculation, the Na^+-taurocholate co-transporting polypeptide (NTCP; also known as SLC10A1) expressed in hepatocytes, and the apical sodium-dependent bile acid transporter (ASBT; also known as SLC10A2) expressed on enterocytes in the terminal ileum[2]. In recent years, ASBT has attracted much interest as a potential drug target for treatment of hypercholesterolaemia, because inhibition of ASBT reduces reabsorption of bile acids, thus increasing bile acid synthesis and consequently cholesterol consumption[3,4]. However, a lack of three-dimensional structures of bile acid transporters hampers our ability to understand the molecular mechanisms of substrate selectivity and transport, and to interpret the wealth of existing functional data[2,5–8]. The crystal structure of an ASBT homologue from *Neisseria meningitidis* ($ASBT_{NM}$) in detergent was reported recently[9], showing the protein in an inward-open conformation bound to two Na^+ and a taurocholic acid. However, the structural changes that bring bile acid and Na^+ across the membrane are difficult to infer from a single structure. To understand the structural changes associated with the coupled transport of Na^+ and bile acids, here we solved two structures of an ASBT homologue from *Yersinia frederiksenii* ($ASBT_{Yf}$) in a lipid environment, which reveal that a large rigid-body rotation of a substrate-binding domain gives the conserved 'crossover' region, where two discontinuous helices cross each other, alternating accessibility from either side of the cell membrane. This result has implications for the location and orientation of the bile acid during transport, as well as for the translocation pathway for Na^+.

Purified $ASBT_{Yf}$, when reconstituted into liposomes, mediates Na^+-dependent transport of the conjugated bile acid taurocholic acid (TCA) with an apparent Michaelis constant (K_m) of 46.8 ± 7.4 μM (Fig. 1a, b and Extended Data Fig. 1). $ASBT_{Yf}$ was crystallized in lipidic cubic phase (LCP), and the structure solved to 1.95 Å (Extended Data Table 1). $ASBT_{Yf}$ has ten transmembrane segments (TM1–10) divided into two domains: a panel domain, formed by TM1, 2, 6 and 7; and a core domain, formed by TM3–5 and 8–10. The first and last five transmembrane helices are structurally homologous, and due to their respective inverted topology give $ASBT_{Yf}$ an internal two-fold pseudosymmetry axis (Fig. 1c and Extended Data Fig. 2). In the core domain, TM4 and 9 unwind in the middle of the membrane and cross each other (Fig. 1d), a structural motif also observed in $ASBT_{NM}$. In addition to the transmembrane helices, there are four amphipathic helices, AH1–4 (Extended Data Fig. 2), that are probably located at the interface between the membrane and the bulk solution, and can be used to infer the approximate position of the lipid bilayer. $ASBT_{Yf}$ assumes an inward-open conformation in which the panel and the core domains contact at the extracellular side, creating a large cavity solvent accessible only from the cytoplasm that extends as far as the crossover region (Fig. 1d).

In the structure of $ASBT_{NM}$, which has roughly 40% sequence identity with $ASBT_{Yf}$, two Na^+-binding sites were identified, Na1 and Na2, which are both located in the core domain behind the crossover (Fig. 1d and Extended Data Fig. 3a, b). Na2 sits directly between the carboxy (C)-terminal ends of helices TM4a and TM9a; Na1 is positioned roughly 8 Å away between TM4b, TM9a and TM5. The residues coordinating Na^+ are highly conserved between $ASBT_{NM}$ and $ASBT_{Yf}$ (Extended Data Fig. 4), but there is no obvious electron density at Na1 in the $ASBT_{Yf}$ structure that could be attributed to Na^+, and a very weak density at Na2 (Extended Data Fig. 3c, d). A closer examination of the residues forming the putative Na^+-binding sites in $ASBT_{Yf}$ showed that they are not in a position to coordinate Na^+ optimally, probably due to a conformational change of TM4b. Whereas other transmembrane helices in the core domain of $ASBT_{Yf}$ align closely with those of $ASBT_{NM}$, TM4b tilts $\sim 11^\circ$ away from the crossover, and its first helix turn unwinds (Fig. 1e). These changes bring Asn 109 and Ser 108 out of range for coordination of Na^+ in Na1, and may also affect the orientation of backbone carbonyls that form part of Na2 (Fig. 1d and Extended Data Fig. 5a, b). To test whether $ASBT_{Yf}$ contains two Na^+-binding sites, like $ASBT_{NM}$, we measured $^{22}Na^+$ binding by purified $ASBT_{Yf}$ (Fig. 1f). Wild-type $ASBT_{Yf}$ bound $^{22}Na^+$ with an apparent half-maximum effective concentration (EC_{50}) of 5.37 ± 0.01 mM and a Hill coefficient of 1.56 ± 0.06, suggesting cooperative binding between more than one Na^+-binding site. Consistent with the notion that Na1 and Na2 are two Na^+-binding sites in $ASBT_{Yf}$, replacing Glu 254 in Na1 or Gln 258 in Na2 with Ala reduced binding of $^{22}Na^+$ to 49% and 68% when compared with wild-type $ASBT_{Yf}$, respectively, and reduced the Hill coefficients to 1.06 ± 0.02 and 0.5 ± 0.1. The structure of $ASBT_{Yf}$ thus represents an inward-facing unliganded state lacking Na^+ and bile acid. Interestingly, the rotation of TM4b also renders Na1 accessible to the solvent from the intracellular side (Extended Data Fig. 5c, d), presenting a potential pathway for release of Na^+ into the cytosol.

To obtain $ASBT_{Yf}$ in an alternative conformation, we perturbed Na1 by mutating the highly conserved Glu 254 to Ala. Although $ASBT_{Yf}$(E254A) is still capable of mediating Na^+-dependent transport (Extended Data Fig. 1c–e), the rate of TCA uptake is substantially reduced. Like wild-type $ASBT_{Yf}$, the E254A mutant was crystallized in LCP, and a complete data set was collected to 2.5 Å resolution. Interestingly, molecular replacement using the full structure of wild-type $ASBT_{Yf}$ as a search model did not yield a valid solution. However, when the panel and core domain were used as two independent rigid bodies, a single solution was obtained (Fig. 2a and Extended Data Table 1). The core and panel domains from $ASBT_{Yf}$(E254A) individually align well with those of the wild type, with α-carbon root mean squared deviation (r.m.s.d.)

[1]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas 77030, USA. [2]Department of Physiology and Cellular Biophysics, Columbia University, New York, New York 10032, USA. [3]New York Consortium on Membrane Protein Structure, New York, New York 10027, USA. [4]Department of Psychiatry and Center for Molecular Recognition, Columbia University, New York, New York 10032, USA. [5]New York State Psychiatric Institute, Division of Molecular Therapeutics, New York, New York 10032, USA. [6]Ion Channel Research and Drug Development Center, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China.
*These authors contributed equally to this work.

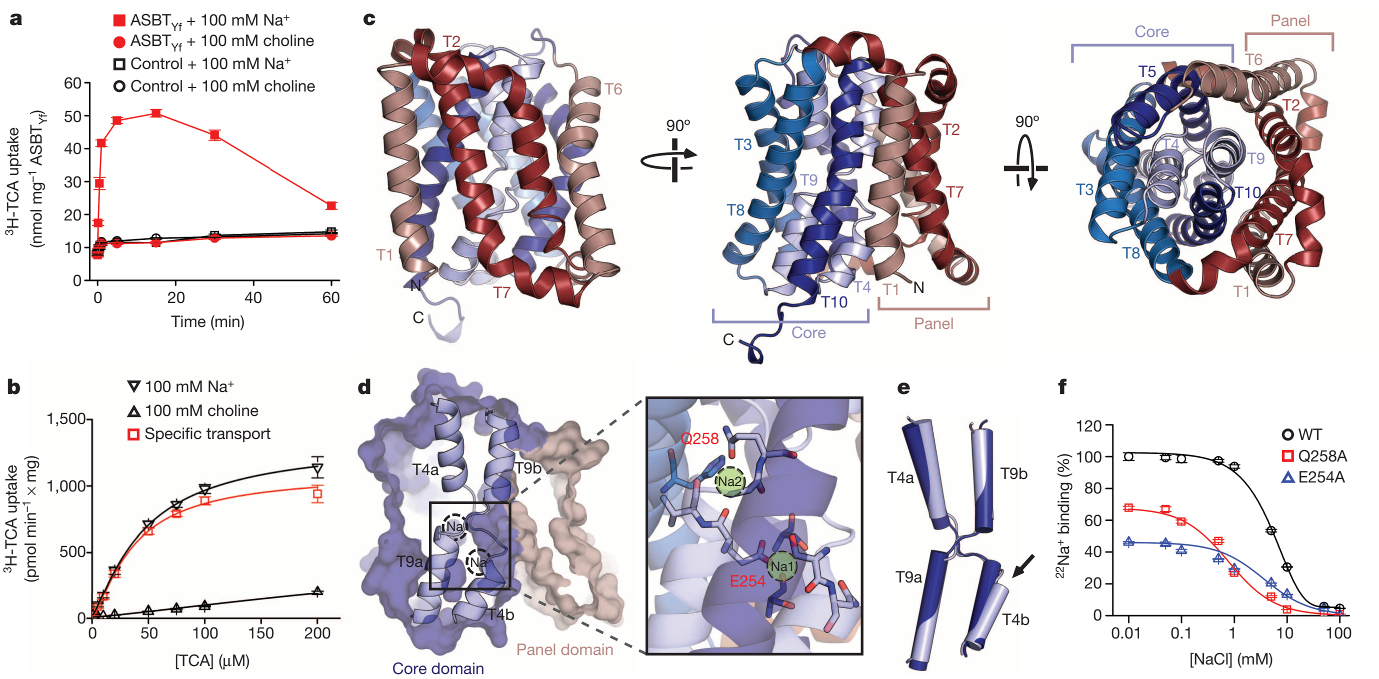


Figure 1 | Function and crystal structure of ASBT$_{Yf}$. **a**, Time course of uptake of ^{3}H-TCA into empty or ASBT$_{Yf}$-containing proteoliposomes in the presence of 100 mM external NaCl or choline chloride. **b**, Uptake of ^{3}H-TCA 30 s after addition to ASBT$_{Yf}$-containing proteoliposomes in the presence of 100 mM external NaCl or choline chloride, as a function of the external ^{3}H-TCA concentration. **c**, Cartoon representation of the ASBT$_{Yf}$ structure shown from two perpendicular directions in the plane of the membrane with the periplasm on top (left and middle), and from the extracellular side (right). The transmembrane helices are coloured in pseudosymmetry-related pairs. **d**, A cutaway surface representation of ASBT$_{Yf}$ showing the locations of the discontinuous helices TM4 and TM9 relative to the intracellular cavity. Locations of the Na$^+$-binding sites in the previously reported ASBT$_{NM}$ structure are marked with dotted circles. Inset on the right shows a magnified view of residues that coordinate Na$^+$ in the ASBT$_{NM}$ structure. **e**, Alignment of the discontinuous helices TM4 and TM9 in the ASBT$_{Yf}$ (light blue) and ASBT$_{NM}$ (dark blue) structures. The partly unwound region of TM4b is marked with a black arrow. **f**, Na$^+$-binding kinetics of ASBT$_{Yf}$ and the Na$^+$-site mutants. Equilibrium binding of 0.95 μM [^{22}Na]Cl (5.92 Ci mmol^{-1}) to 250 ng of wild-type (WT), E254A or Q258A ASBT$_{Yf}$ was measured with the SPA in the presence of increasing NaCl concentrations ranging from 0–100 mM. Isotopic replacement of ^{22}Na$^+$ was plotted as a function of the concentration of non-labelled NaCl. The means of triplicate measurements ± standard error of the mean (s.e.m.) were subjected to nonlinear regression fitting in Prism 5 (GraphPad) for panels **a**, **b** and **f**.

values of 0.9 and 1.7 Å respectively, compared with an r.m.s.d. of 3.7 Å after aligning both domains. This indicates that there is relative motion between the core and panel domains. As the amphipathic helices in the panel domain probably remain at the membrane–solvent interfaces (Extended Data Fig. 6), the core domain in the E254A mutant must undergo a rigid-body rotation, causing residues on TM4b and TM9b lining the interface with the panel domain to translate 6–9 Å towards the periplasm (Fig. 2a and Supplementary Video 1). This motion is facilitated by small changes in AH2, AH4 and the TM5–6 loop, which act as hinges; a kink also forms in TM1 at Pro 10. Whereas in the wild-type structure, the cavity between the two domains is sealed on the extracellular side by interactions between TM9b, TM2 and TM7, in the E254A structure the two domains now form contacts between TM4b, TM2 and TM7 at the intracellular side. This creates a deep cavity allowing solvent access to the crossover region from the extracellular side (Fig. 2b). Similarly to the wild-type ASBT$_{Yf}$ structure, the E254A structure is not bound to Na$^+$ or a bile acid, and thus, the structure of E254A is probably in the outward-open unliganded conformation.

To address the question of how this rigid-body motion of the core domain can translocate bile acid across the membrane, we compared the solvent-accessible surfaces of the inward- and outward-facing cavities in the two structures. This analysis reveals a narrow area running across the centre of the core and panel domains that is accessible to the solvent in both the inward-open and outward-open conformations (Fig. 2c). The dual-accessibility region includes the crossover region, and contains residues that are highly conserved among ASBT homologues. To test whether the crossover region is indeed accessible from the periplasm, as predicted by the E254A structure, and that the outward-open conformation is not simply an artefact caused by detergent solubilization and/or mutation to Na1, we measured the accessibility of introduced cysteines in the domain interface to modification by a membrane-impermeable pegylating reagent either from the periplasm of intact *Escherichia coli* cells, or from both sides of the membrane in cells ruptured by sonication (Fig. 2d, e). The cysteine mutants were still able to transport TCA, albeit at a reduced rate, indicating that they probably undergo similar conformational changes as wild-type ASBT$_{Yf}$ (Extended Data Fig. 7a). All three mutants were pegylated when exposed to the reagent in ruptured cells, but the cysteine at position 123, which is accessible only from the intracellular side in both conformations, was not pegylated in intact cells. Despite being inaccessible to the periplasm in the wild-type structure, the T106C mutation located near the crossover was pegylated in intact cells, suggesting that the crossover region is indeed accessible from the extracellular side, as observed in the E254A structure.

In the ASBT$_{NM}$ structure, a TCA was built into the inward-facing cavity with an orientation roughly perpendicular to the membrane, that is, with the cholesterol ring close to the crossover and the taurine group extending to the intracellular entrance of the cavity. Curiously, most of the residues forming the TCA-binding site in the ASBT$_{NM}$ structure are accessible only in the inward-open state (Fig. 2c). In the outward-open state, this TCA-binding site is buried in the protein matrix without an access pathway from the periplasm. This observation seems at odds with the alternating-access model[10,11], according to which ligand is transported across the membrane via the alternating exposure of central ligand-binding sites to the intracellular or the extracellular space through a sequence of distinct conformational alterations. Structural validation of the principles of the alternating-access mechanism has been obtained from crystal structures of several transporters[12–20]. One possible resolution

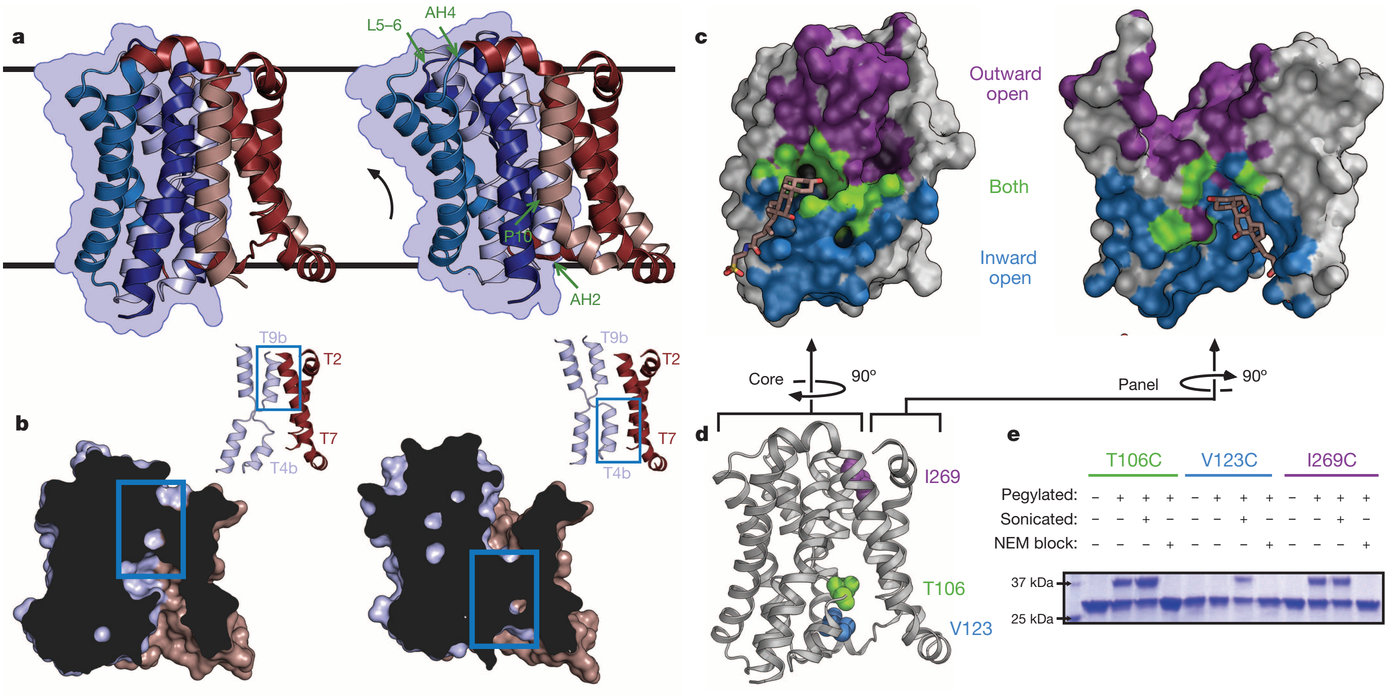


Figure 2 | Structure and validation of the outward-open conformation of ASBT$_{Yf}$. **a**, The wild-type (left) and E254A (right) structures are shown side by side. Black lines correspond to the approximate position of the lipid bilayer inferred from the amphipathic helices. The core domain is marked with a blue silhouette, and regions acting as hinges in the conformational change between the two structures are marked with green arrows in the E254A structure. **b**, Cutaway view of the surfaces of the wild-type (left) and E254A (right) structures, showing the intracellular and extracellular cavities. Insets show key helices forming the interface between the panel and core domain; blue rectangles show the location of the interdomain interface in both structures. **c**, Surface representations of the core (left) and panel (right) domains of the wild-type ASBT$_{Yf}$ structure. The sides of both domains facing the central cavity are coloured according to whether they are accessible to the cytoplasm in the wild-type structure (blue), accessible to the periplasm in the E254A structure (violet), or are solvent accessible in both conformations (green). A stick representation of TCA marks the location of the substrate in the ASBT$_{NM}$ structure. **d**, Cartoon representation of the wild-type ASBT$_{Yf}$ structure with the locations of Thr 106, Val 123 and Ile 269 marked with spheres. **e**, SDS–polyacrylamide gel electrophoresis (SDS–PAGE) gel showing results of pegylation experiments for the three ASBT$_{Yf}$ cysteine mutants. Pluses and minuses mark whether or not samples were incubated with mPEG-Mal-5K, were sonicated before pegylation to rupture the cell membranes, or were incubated with *N*-ethylmaleimide (NEM) before pegylation to prevent further modification of the cysteine residues.

for this apparent conflict is that the transporter possesses another as yet unobserved binding site for TCA, which, unlike the binding site shown in ASBT$_{NM}$, has alternating access to both the periplasm and cytoplasm. For example, the bile acid could bind in a lateral orientation to the dual-access region described earlier (Extended Data Fig. 7c), with its mostly hydrophobic β-face oriented towards the hydrophobic panel domain, and its hydrophilic α-face close to polar residues on the core domain (Extended Data Fig. 7d). In this configuration, the rigid-body motion of the core domain revealed by the two ASBT$_{Yf}$ structures would be sufficient to translocate the bile acid across the membrane. Mutation of polar residues capable of forming hydrogen bonds to the three hydroxyls on the steroid nucleus of TCA in a speculative horizontal orientation reduces TCA binding relative to wild-type ASBT$_{Yf}$ (Extended Data Fig. 7e, f). However, introducing mutations to the protein can affect substrate binding in detergent indirectly through a variety of mechanisms not involving direct contact with the substrate, and further experimental validation will be required to demonstrate the existence of a horizontal binding site in ASBT.

Experimental observation of distinct conformations of secondary transporters during the transport cycle is a major challenge in understanding the transport-associated dynamics of these molecular machines. We have presented two alternative conformations of a transporter, produced by a point mutation in a Na^+ site. It cannot be fully excluded that mutating Na1 resulted in perturbations from the native structure, but the two structures appear to correspond to ligand-free inward- and outward-open states. To investigate further the states of the ASBT transport cycle, we measured the interdependency of Na^+ and TCA binding with the scintillation proximity assay (SPA). Binding of TCA to ASBT$_{Yf}$ is strongly dependent on the concentration of Na^+ (Fig. 3a), whereas TCA has a minimal effect on Na^+ binding (Fig. 3b). This suggests that the Na^+ sites are occupied before TCA can bind to the transporter. From these results, we can begin to enumerate and order conformational states in a preliminary model of the ASBT transport cycle (Fig. 3c).

ASBT$_{Yf}$ and ASBT$_{NM}$ share their fold with NhaA (a member of the Na^+/H^+ antiporter family). NhaA also possesses inverted psuedosymmetry repeats that form a substrate-binding core domain and a panel domain, although the panel domain typically contains two or more additional helices that form a homodimer interface[21] (Extended Data Fig. 8a). Two Na^+/H^+ antiporter structures (NhaA and NapA) have been published[21,22], which correspond to inward-open and outward-open states. Despite the substantial difference in the size of the substrates involved, the rigid-body movement of the core domain that converts between the two states in the Na^+/H^+ antiporters is remarkably similar to that observed for ASBT$_{Yf}$ (Extended Data Fig. 8b, c). As in ASBT$_{Yf}$, the conformational change provides alternating access to the crossover region, where Na^+ and protons are predicted to bind to a cluster of conserved acidic residues. However, this site is not equivalent to either Na1 or Na2 in the bile acid transporter. Examination of the site on NapA corresponding to Na2 shows that two of the polar residues coordinating Na^+ in ASBT$_{Yf}$ (Gln 258 and His 71; Extended Data Fig. 8f) are in fact replaced with two positively charged side chains (Arg 331 and Lys 305; Extended Data Fig. 8g), which form hydrogen bonds with the C-terminal ends of helices TM4a and TM11a (TM4a and TM9a in ASBT$_{Yf}$). In the structurally unrelated antiporter CaiT, which is Na^+ independent, an arginine residue has recently been shown to mimic binding of Na^+ to a site found in Na^+-dependent symporters of the same fold[23]. It might therefore be

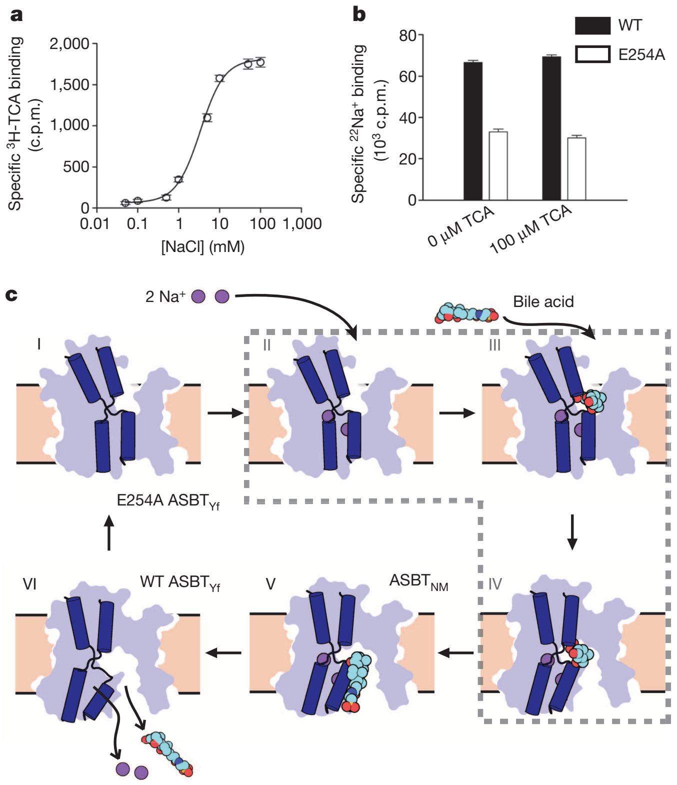


Figure 3 | Proposed ASBT$_{Yf}$ transport mechanism. **a**, Binding of ^{3}H-TCA to detergent-solubilized wild-type ASBT$_{Yf}$ as a function of NaCl concentration, as measured by SPA. c.p.m., counts per minute. **b**, Binding of ^{22}Na$^+$ to detergent-solubilized wild-type (WT) and E254A ASBT$_{Yf}$ in the presence and absence of 100 μM TCA. Error bars are the s.e.m. of triplicate measurements. **c**, Key conformational states of ASBT during the translocation of substrates. Distinct conformations captured by crystallography are indicated with the name of the relevant protein, whereas hypothetical structural states are surrounded with a dashed grey line. In the ligand-free state (I), corresponding to the E254A ASBT$_{Yf}$ structure, the crossover region is exposed to the periplasm. Na$^+$ then binds to Na1 and Na2 (II), facilitating the binding of TCA, probably to the dual-accessibility region (III). Conversion to the inward-open conformation (IV) allows TCA access to the binding site observed in the ASBT$_{NM}$ structure (V). Exposure to lower Na$^+$ concentrations in the cytoplasm drives release of Na$^+$, possibly by the pathway opened by the rotation of TM4b in the ASBT$_{Yf}$ structure (VI), which in turn triggers release of TCA.

feasible to speculate that these residues in NapA can have a role analogous to bound Na$^+$ in ASBT$_{Yf}$ and ASBT$_{NM}$. Further comparison of these two families of transporters may provide insight into how the same fold and conformational change can act as 'scaffolding' for highly distinct substrates and coupling mechanisms. The ASBT$_{Yf}$ structure also shows some interesting parallels with structurally unrelated families of transporters. For example, the translation towards the periplasm observed in ASBT$_{Yf}$(E254A) is similar to the elevator-like motion of the substrate-binding domain in Glt$_{Ph}$, a homologue of the glutamate transporter[17,24] (Extended Data Fig. 8d).

Although they belong to an unrelated fold, the structures of ASBT$_{Yf}$ invite comparison to recent structures obtained for *E. coli* XylE, a xylose/H$^+$ symporter belonging to the major facilitator superfamily (MFS) with homology to the mammalian GLUT transporters. Like the bile acid transporters, MFS transporters contain two domains, with a substrate-binding site located at the domain interface, and also possess inverted pseudosymmetry repeats. However, whereas in the bile acid transporter fold helices from each repeat are interleaved between the asymmetric panel and core domains, in the MFS fold each of the two pseudosymmetry repeats forms a separate six-helix domain. Recently, structures of XylE from *E. coli* have been solved in the inward- and outward-facing conformations, as well as a potential intermediate conformation[15,18] (Extended Data Fig. 8e). Comparison of the three states shows that, like ASBT$_{Yf}$, XylE provides alternating access to the central substrate-binding site by the rigid-body motion of a mobile domain relative to a fixed domain, although in XylE the conformational change is largely a rocking motion that does not translate the binding site towards the opposite side of the bilayer.

In XylE, salt bridges between conserved residues in the two domains have been proposed to have a key role in conversion between the inward- and outward-open conformations[15]. In contrast, in ASBT$_{Yf}$ there are few hydrophilic interactions between polar or charged residues on the core domain and the very hydrophobic panel domain, a feature that may lower the energy barrier to moving the two domains relative to each other. Regardless, in both structural folds, even with multiple conformations of the same transporter, it is still not obvious how binding of Na$^+$ or H$^+$ and the cognate substrate could trigger a conversion between the two alternate conformations. Further studies are necessary to understand how the energy from Na$^+$ or H$^+$ binding triggers and drives the conformational changes required for binding and translocation of the cognate substrate.

In humans, ASBT inhibitors have received considerable attention as potential therapeutics for the treatment of hypercholesterolaemia[25] and type 2 diabetes[26]. Another possible medical application of compounds targeting bile acid transporters involves conjugating bile acids to drugs with poor oral bioavailability, so that they are recognized as substrates by ASBT and NTCP and absorbed in the intestine and liver[27]. Both approaches would greatly benefit from an improved understanding of bile acid transporter structure and mechanism of action. ASBT$_{Yf}$ shares 22% sequence identity and 59% similarity with human ASBT. Additionally, the residues forming the two Na$^+$-binding sites are highly conserved (Extended Data Fig. 4). This suggests that the overall fold and transport mechanism are similar between the two proteins, and that ASBT$_{Yf}$ may serve as a useful model system for understanding mechanisms of transport and inhibition in the mammalian ASBT homologues.

METHODS SUMMARY

The gene encoding ASBT$_{Yf}$ (RefSeq accession ZP_04633709.1) was obtained by PCR from the genomic DNA of *Y. frederiksenii*, ligated into a modified pET vector, and expressed in *E. coli*. The ASBT$_{Yf}$ protein was extracted from the cell membrane, purified by a metal affinity column, and further purified by size-exclusion chromatography. Crystals were grown in LCP with 30% (v/v) PEG-400, 0.1 M Na-citrate pH 5.5, 0.1 M NaCl and 3% (w/v) D-trehalose for the wild-type protein and 39% (v/v) PEG-400, 0.1 M Tris-HCl pH 8.5, 0.1 M KCl and 10 mM $MnCl_2$ for the E254A mutant. The wild-type structure was solved by molecular replacement using the structure of ASBT$_{NM}$ (PDB accession 3ZUX) as a search model; the E254A mutant was solved by molecular replacement with the separate core and panel domains of the wild-type structure. All structure figures were made in Pymol. ^{22}Na$^+$ and ^{3}H-TCA binding were measured with the SPA[28,29], and the solvent accessibility of ASBT$_{Yf}$ cysteine mutants was measured by modification with mPEG-Mal-5K. Uptake of ^{3}H-TCA into ASBT$_{Yf}$-containing proteoliposomes of polar *E. coli* lipid extracts (Avanti) was measured with a rapid filtration assay as described previously[30].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 March; accepted 24 October 2013.
Published online 8 December 2013.

1. Dawson, P. A. Role of the intestinal bile acid transporters in bile acid and drug disposition. *Handb. Exp. Pharmacol.* **201,** 169–203 (2011).
2. Claro da Silva, T., Polli, J. E. & Swaan, P. W. The solute carrier family 10 (SLC10): beyond bile acid transport. *Mol. Aspects Med.* **34,** 252–269 (2013).
3. West, K. L., Ramjiganesh, T., Roy, S., Keller, B. T. & Fernandez, M. L. 1-[4-[4[(4*R*,5*R*)-3,3-Dibutyl-7-(dimethylamino)-2,3,4,5-tetrahydro-4-hydroxy-1,1-di oxido-1-benzothiepin-5-yl]phenoxy]butyl]-4-aza-1-azoniabicyclo[2.2.2]octane methanesulfonate (SC-435), an ileal apical sodium-codependent bile acid transporter inhibitor alters hepatic cholesterol metabolism and lowers plasma low-density lipoprotein-cholesterol concentrations in guinea pigs. *J. Pharmacol. Exp. Ther.* **303,** 293–299 (2002).
4. Braun, A. *et al.* Inhibition of intestinal absorption of cholesterol by ezetimibe or bile acids by SC-435 alters lipoprotein metabolism and extends the lifespan of SR-BI/apoE double knockout mice. *Atherosclerosis* **198,** 77–84 (2008).

5. Hagenbuch, B., Stieger, B., Foguet, M., Lubbert, H. & Meier, P. J. Functional expression cloning and characterization of the hepatocyte Na^+/bile acid cotransport system. *Proc. Natl Acad. Sci. USA* **88,** 10629–10633 (1991).
6. Wong, M. H., Oelkers, P., Craddock, A. L. & Dawson, P. A. Expression cloning and characterization of the hamster ileal sodium-dependent bile acid transporter. *J. Biol. Chem.* **269,** 1340–1347 (1994).
7. Alrefai, W. A. & Gill, R. K. Bile acid transporters: structure, function, regulation and pathophysiological implications. *Pharm. Res.* **24,** 1803–1823 (2007).
8. Doring, B., Lutteke, T., Geyer, J. & Petzinger, E. The SLC10 carrier family: transport functions and molecular structure. *Curr. Top. Membr.* **70,** 105–168 (2012).
9. Hu, N. J., Iwata, S., Cameron, A. D. & Drew, D. Crystal structure of a bacterial homologue of the bile acid sodium symporter ASBT. *Nature* **478,** 408–411 (2011).
10. Jardetzky, O. Simple allosteric model for membrane pumps. *Nature* **211,** 969–970 (1966).
11. Widdas, W. F. Inability of diffusion to account for placental glucose transfer in the sheep and consideration of the kinetics of a possible carrier transfer. *J. Physiol. (Lond.)* **118,** 23–39 (1952).
12. Dang, S. *et al.* Structure of a fucose transporter in an outward-open conformation. *Nature* **467,** 734–738 (2010).
13. Huang, Y., Lemieux, M. J., Song, J., Auer, M. & Wang, D. N. Structure and mechanism of the glycerol-3-phosphate transporter from *Escherichia coli*. *Science* **301,** 616–620 (2003).
14. Krishnamurthy, H. & Gouaux, E. X-ray structures of LeuT in substrate-free outward-open and apo inward-open states. *Nature* **481,** 469–474 (2012).
15. Quistgaard, E. M., Low, C., Moberg, P., Tresaugues, L. & Nordlund, P. Structural basis for substrate transport in the GLUT-homology family of monosaccharide transporters. *Nature Struct. Mol. Biol.* **20,** 766–768 (2013).
16. Ressl, S., Terwisscha van Scheltinga, A. C., Vonrhein, C., Ott, V. & Ziegler, C. Molecular basis of transport and regulation in the Na^+/betaine symporter BetP. *Nature* **458,** 47–52 (2009).
17. Reyes, N., Ginter, C. & Boudker, O. Transport mechanism of a bacterial homologue of glutamate transporters. *Nature* **462,** 880–885 (2009).
18. Sun, L. *et al.* Crystal structure of a bacterial homologue of glucose transporters GLUT1–4. *Nature* **490,** 361–366 (2012).
19. Yamashita, A., Singh, S. K., Kawate, T., Jin, Y. & Gouaux, E. Crystal structure of a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporters. *Nature* **437,** 215–223 (2005).
20. Yernool, D., Boudker, O., Jin, Y. & Gouaux, E. Structure of a glutamate transporter homologue from *Pyrococcus horikoshii*. *Nature* **431,** 811–818 (2004).
21. Hunte, C. *et al.* Structure of a Na^+/H^+ antiporter and insights into mechanism of action and regulation by pH. *Nature* **435,** 1197–1202 (2005).
22. Lee, C. *et al.* A two-domain elevator mechanism for sodium/proton antiport. *Nature* **501,** 573–577 (2013).
23. Kalayil, S., Schulze, S. & Kuhlbrandt, W. Arginine oscillation explains Na^+ independence in the substrate/product antiporter CaiT. *Proc. Natl Acad. Sci. USA* **110,** 17296–17301 (2013).
24. Jensen, S., Guskov, A., Rempel, S., Hänelt, I. & Slotboom, D. J. Crystal structure of a substrate-free aspartate transporter. *Nature Struct. Mol. Biol.* **20,** 1224–1226 (2013).
25. Kramer, W. & Glombik, H. Bile acid reabsorption inhibitors (BARI): novel hypolipidemic drugs. *Curr. Med. Chem.* **13,** 997–1016 (2006).
26. Chen, L. *et al.* Inhibition of apical sodium-dependent bile acid transporter as a novel treatment for diabetes. *Am. J. Physiol. Endocrinol. Metab.* **302,** E68–E76 (2012).
27. Tolle-Sander, S., Lentz, K. A., Maeda, D. Y., Coop, A. & Polli, J. E. Increased acyclovir oral bioavailability via a bile acid conjugate. *Mol. Pharm.* **1,** 40–48 (2004).
28. Shi, L., Quick, M., Zhao, Y., Weinstein, H. & Javitch, J. A. The mechanism of a neurotransmitter:sodium symporter—inward release of Na^+ and substrate is triggered by substrate in a second binding site. *Mol. Cell* **30,** 667–677 (2008).
29. Levin, E. J., Quick, M. & Zhou, M. Crystal structure of a bacterial homologue of the kidney urea transporter. *Nature* **462,** 757–761 (2009).
30. Quick, M. & Javitch, J. A. Monitoring the function of membrane transport proteins in detergent-solubilized form. *Proc. Natl Acad. Sci. USA* **104,** 3603–3608 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements Data for this study were collected at beamlines 8.2.2 at the Advanced Light Source, X-29 at the National Synchrotron Light Source, and 24ID-E and 17ID-B at the Advanced Photon Source. This work was supported by the US National Institutes of Health (R01DK088057, R01GM098878, U54GM095315 and U54GM087519), the American Heart Association (12EIA8850017), and the Cancer Prevention and Research Institute of Texas (R12MZ). M.Z. thanks R. MacKinnon for advice and guidance on scientific directions.

Author Contributions X.Z., E.J.L., M.Q. and M.Z. conceived the project and designed the research. X.Z., E.J.L., M.Q., Y.P., J.G.M., R.S., B.K., R.B. and M.Z. performed experiments. E.J.L. and M.Z. wrote the manuscript with input from all authors.

Author Information Atomic coordinates and structure factors have been deposited at the Protein Data Bank under accessions 4N7W and 4N7X. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.Z. (mzhou@bcm.edu or mingzhou@mail.kiz.ac.cn) or M.Q. (mq2102@columbia.edu).

CORRIGENDUM

doi:10.1038/nature12906

Corrigendum: Deglacial pulses of deep-ocean silicate into the subtropical North Atlantic Ocean

A. N. Meckler, D. M. Sigman, K. A. Gibson, R. François, A. Martínez-García, S. L. Jaccard, U. Röhl, L. C. Peterson, R. Tiedemann & G. H. Haug

Nature **495,** 495–498 (2013); doi:10.1038/nature12006

In Figure 3c of our Letter, the time axis of the ln(Si/Al) record of ODP Site 658A (orange line) was based on a prior age model that did not take into account an age tie-point at 17.8 kyr ago, derived from oxygen isotope stratigraphy (see Methods). This incorrectly shifted the plotted data between 20 kyr ago and 14.2 kyr ago towards younger ages. The shift is most significant at the onset of the opal peak during Heinrich stadial 1 (HS1), which in the original Fig. 3 appeared to occur at 16.5 kyr ago, whereas the correct age model puts this increase in ln(Si/Al) at 17.8 kyr ago. Figure 2, which shows the full time series, is not affected. The error has no implications for the conclusions of the original Letter. In fact, the earlier onset of the opal peak in the correct age model is even more consistent with our interpretation of the opal peak as an indication of changes in ocean circulation early during deglaciations. The correct age model and data are accessible at Pangaea (http://doi.pangaea.de/10.1594/PANGAEA.810016). Figure 3 has been corrected in the PDF and HTML of the original paper.

CAREERS

TURNING POINT Former dancer gives up the barre to study narwhals in the Arctic **p.577**

WORKPLACE Regular exercise helps people feel better about work–life balance **p.577**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

CARGO/IMAGEZOO/CORBIS

CHARITIES

Profiting from non-profits

Scientists interested in doling out funds to worthy grant recipients may thrive working at science foundations.

BY ALLA KATSNELSON

Anita Pepper's second career, as administrator of biomedical grant programmes at a non-profit organization, came about through a combination of luck and circumstance. A geneticist by training, she was approaching the end of a five-year postdoctoral fellowship at the University of Pennsylvania in Philadelphia in 2008 and was thinking about what to do next. She and her husband did not want to relocate, but she knew that an academic job search in a single city was too limited to be successful. And although she enjoyed bench work, she was a naturally social person who liked the idea of interacting more frequently with people.

So Pepper set out on a mission to meet anyone in Philadelphia who held leadership positions, particularly those involved in improving health through education or research. Lunch with a brand executive — who worked with companies and others to design brand identities — led her to another contact who told her about a senior position at the Pew Charitable Trusts in Philadelphia and Washington DC. Pew wanted someone to help administer two biomedical grant initiatives, the Pew Latin American Fellows Program and the Pew Scholars Program. Pepper eagerly applied, captivated by the opportunity to help young scientists to find their footing in an increasingly challenging climate — a struggle she could relate to. Pew hired her as a senior associate in 2008. "It was completely random," she says. "I didn't even know that these jobs existed." She was promoted to director of the programmes in 2012.

CHARITIES CALLING

For people more attracted to facilitating research than to toiling at the bench themselves, working for a foundation or charity could be a great fit. Programme officer jobs in such organizations offer healthy salaries — starting in the range of US$60,000–110,000 in the United States and €30,000–80,000 (US$41,000–109,000) in Europe. Yet many, like Pepper, simply do not know about the existence of such jobs. Even those who do often have mistaken ideas about what they entail. "I think many people have the impression that you're just reading grant applications," says Michael Madeja, managing director at the Hertie Foundation in Frankfurt, the largest private funder of brain research in Germany. "But it's a job with travelling and meeting people, a lot of communication, ▶

and developing your own agenda — it's a very creative situation," says Madeja, who also maintains a lab at the University of Frankfurt.

Most people who escape from a lab to a foundation or other private funder start out as a programme associate or director, overseeing grant programmes that further their organization's agenda by supporting a general area of research, funding work on a specific topic or in a region, or helping a particular group such as early-career scientists. And yes — reading grant applications is almost always part of the gig.

HERTIE FOUNDATION

"What we are looking for is very intelligent, engaged people."
Michael Madeja

"If you don't like reviewing, this is not a job for you," says Richard Wiener, a programme director at the Research Corporation for Science Advancement, a foundation based in Tucson, Arizona, that supports the physical sciences. But just as important is building relationships with current and prospective grant recipients and the research community. "You really need to know your science, but you're also really focused on connecting and developing people, and establishing new collaborations and opportunities," Wiener says.

Communication skills are thus at the heart of such a career. Developing a deep understanding of the topic or population that one's organization funds involves attending and perhaps arranging conferences and cultivating contacts with established experts, up-and-coming researchers and scientists who a programme already funds. Many foundations work closely with grant or fellowship applicants to help them to strengthen their proposals. Some, like Wiener's, also spend considerable time advising researchers whose applications were denied. "It's a supportive role," he says.

The programme manager's precise duties depend on the types of project, but variety is often a key feature. Marta Tufet, an international-activities adviser for science funding at the Wellcome Trust in London, the United Kingdom's largest biomedical research charity, is part of a six-person team that develops the strategy and policy for the Trust's scientific research portfolio in low- and middle-income countries. The team has, for example, funded projects across 51 institutions in Africa, aiming to strengthen universities' research capacity. She seeks out deserving researchers, advises them through the application process and helps to monitor their progress once a programme has been funded. Tufet travels frequently to the African institutions, attending scientific advisory-board and steering-committee meetings and advising on operation and governance matters, as well as on institutions' scientific directions. "It allows me, early on in my career, to have quite a lot of input at a high level, which you wouldn't necessarily have as a young researcher," she says.

Non-profit funding organizations vary greatly in size, structure and culture, so it is helpful for early-career scientists to think about what interests them most. A larger organization might have more of an impact on a field, even if only in the amount of money granted, whereas a smaller one may offer its staff a wider array of responsibilities and more flexibility, but pay a smaller salary. Some organizations engage exclusively in funding; others have advocacy or education components. At the Alzheimer's Drug Discovery Foundation in New York, for example, most scientists oversee grant programmes. But the foundation hired Penny Dacks to collect and summarize the literature on proposed prevention strategies for Alzheimer's disease and cognitive decline and communicate this information to the public, as well as to researchers and clinicians, through opinion pieces in medical journals.

Potential applicants should also decide whether they are committed to bolstering a particular field of research — perhaps one close to their own — or whether they would prefer to work for an organization that funds many areas within and outside science. Betsy Myers, programme director for medical research at the Doris Duke Charitable Foundation in New York, found that she loved her daily interactions with experts in the broad swathe of subjects that her organization funds, including environmental conservation, performing arts and child well-being programmes. "Do your homework and really think about where your skills and creativity match up with the organization's," she says.

ALEX PARLINI

"You have to have the confidence to go to somebody who you are not sure will be helpful."
Anita Pepper

PROMISING FUTURE?

The proportion of research funded by the non-profit sector seems to be growing. In 2011, private philanthropic organizations in the United States spent about $1.3 billion on biomedical research. In 2012, public charities, which raise money from a multitude of sources, spent $1 billion. Non-profit organizations took a hit from the economic downturn, with some funders cutting or freezing programmes; but the sector is bouncing back, with many organizations reporting that once-stalled programmes are running again.

"Foundations are actually looking to hire more PhD holders because they want to drive research," says Gina Agiostratidou, a senior programme officer at the Leona M. and Harry B. Helmsley Charitable Trust in New York, who oversees the research portfolio of a programme on type 1 diabetes. The increasing squeeze on US government funding "makes the role of the foundation even more important now", she adds.

But none of this means that such jobs are easy to land. Although there are many non-profit organizations out there, each hires relatively few PhD holders. Tufet estimates that there are about 600 staff members at the Wellcome Trust, but fewer than 40 in the scientific-funding division. Dacks says that when her team sought to hire an extra scientist last spring, there were 100 applicants for the job.

There is no clear way in, says Myers. "And once you're in, you often have to learn through an apprenticeship model," she says. There are exceptions: the Wellcome Trust and the Hertie Foundation both offer internships of a few weeks or months to give career changers a taste of a foundation track. But potential applicants will often need to do legwork to learn about organizations that align with their interests.

This may entail searching the Internet for non-profit organizations that fund or support a particular topic, strolling through the exhibitor booths at conferences to find representatives from such organizations, seeking them out at a university career office or working through the contacts in one's professional network in search of people who hold related posts. The next step is to request an informal chat to find out about prospects at the organization. "You have to have the confidence to go to somebody who you are not sure will be helpful," says Pepper.

Dacks, who realized in the first year of a three-year postdoc that basic research was not her calling, advises scientists who are on time-limited fellowships to start thinking about their career options early, especially if they suspect that academia might be a poor fit. Another tip is to get involved in non-lab activities — as a student representative, say, on a university or research-society committee — that demonstrate and develop one's abilities beyond the bench.

For the right person, the rewards can be tremendous, says Madeja. Although his organization gives out a relatively modest €10 million per year, he feels that it has made a real difference both to the careers of many talented researchers and in supporting Germany's neuroscience community. He likes the way that his foundation can launch new projects with little red tape and take risks with what it chooses to fund. And he appreciates the family-friendly nature of the workplace. "What we are looking for is very intelligent, engaged people," he says, "who just realize that laboratory work is not for them." ■

Alla Katsnelson *is a freelance writer in Northampton, Massachusetts.*

TURNING POINT
Kristin Laidre

Kristin Laidre was on her way to becoming a dancer until an ankle injury altered her plans. Now a marine biologist at the University of Washington in Seattle, she studies polar bears and narwhals in the Arctic, and has found ways to combine her love of the arts with science.

Did you have an early interest in both dancing and science?
Yes. I grew up with an art-teacher mother in Saratoga Springs, New York, a summer spot for opera singers, ballet dancers and other artists. Surrounded by the arts, I got involved in ballet as a kid and after high school continued performing, eventually with a dance company in Seattle. At the same time, I was a good student who was keen on science, particularly biology. Once I fully accepted that my injury had ended my dancing career, I started pursuing marine-science opportunities in Seattle.

How did you find your first science post?
I visited the US National Oceanic and Atmospheric Administration's National Marine Mammal Lab, found the office of the deputy director and introduced myself. I said I was really interested in a science career working with marine mammals, and wanted to find out how to volunteer. I got lucky. They let me volunteer once a week while working on my undergraduate degree. I learned to catalogue and identify individual humpback whales, do spatial analysis and map beluga whales.

How did you approach science research?
Dancing and science have similar requirements. You have to be super-focused, single-minded and able to spend time alone working for more than 12 hours a day. So I approached it in the same way I did professional dancing.

Did you go straight to graduate studies?
No — I applied to the School of Aquatic and Fishery Sciences at the University of Washington, but was not accepted because of my scores on the Graduate Record Examinations, an admissions requirement for most US graduate schools. I took a year off and retook the exams while doing seal research in Alaska. It taught me to keep going and never give up.

Why did you focus on narwhals for your PhD?
I was interested in Arctic ecology. It was intimidating to accept invitations to be the only woman on a team heading to the Canadian High Arctic for a month to tag narwhals, but when I went, the trip proved inspiring and made me realize that there was much to learn about this difficult-to-study species. This was a turning point that led me to focus my PhD on the ecology of narwhals.

What was the scariest moment?
Nobody teaches you how to run generators, deter polar bears or survive in the Arctic — much less run field surveys — so there is a lot to learn. The scariest moment was definitely when a polar bear picked up the scent of the whale we were tagging and charged into the water towards our boat. Luckily he veered away at the last minute.

You took an artist to the Arctic last year. How did that come about?
Greenland is stunningly beautiful, but data don't capture the essence of all that's changing there. I wrote a grant to the non-profit G. Unger Vetlesen Foundation in New York City, which funds scientific, literary and educational projects, and suggested a three-year project titled 'Imaging the Arctic'. I took artist Maria Coryell-Martin to the field last spring to record the environment through paintings, field sketches and multimedia. We've turned her art and my research into an outreach project that we can take to schools or galleries to teach kids about Arctic ecology and environment.

Is marine biology a competitive field?
That's a hard question to answer. Studying really cool animals is a dream job for many people, but you have to be okay with roughing it. That's not for everyone. And you have to fight for money. Doing a project in the Arctic is really expensive, given the logistics of operating safely when the closest village is 100 kilometres away. There's also the unpredictability of the environment. It does require tenacity. ■

INTERVIEW BY VIRGINIA GEWIN

PHYSICAL SCIENCES

UK recruitment drive

The UK Engineering and Physical Sciences Research Council (EPSRC) is launching doctoral studentships in quantum technologies, robotics and energy and sustainability. Recruitment is under way for 4,400 students as part of the £764-million (US$1.25-billion) scheme. The council announced 3,500 studentships last November and 900 this month, and is spending £390 million to operate 91 new Centres for Doctoral Training across 30 UK universities. Funders include universities and industrial and private partners, where awardees will study and do research. UK research funding has been relatively protected despite austerity policies in recent years, says Lesley Thompson, director of sciences and engineering at the EPSRC.

EU GRANTS

Funding in demand

The European Research Council (ERC) named its first Consolidator Grant recipients on 14 January, all mid-career researchers who are 7–12 years past their PhD. The 312 awardees each received up to €2.75 million (US$3.74 million). The ERC created the new category after the number of mid-stage applicants for its Starting Grant in 2013 jumped to more than 3,600, up by 46% on 2012, reflecting tight funding across Europe. The original Starting Grant is now open to researchers who earned their PhDs 2–7 years ago. The average success rate for both grants is less than 9%. "Competition will remain fierce" for the coveted grants, says ERC president Jean-Pierre Bourguignon.

WORK–LIFE BALANCE

Fit for purpose

Adults who exercise regularly are happiest with their work–life balance, finds a study in the press at *Human Resource Management*. The authors asked 476 US working adults about their exercise behaviour. Those who worked out more than three times a week were most likely to feel positive about managing work and personal duties. Co-author Russell Clayton, who studies management at Saint Leo University in Florida, says that professionals including researchers, whose schedules can preclude long workouts, should aim to fit in brief stints each day. Lab meetings, he suggests, can involve a walk around campus — and scientists should take the stairs. "Find exercise in the margins," he says.

MOTIVATION

Opportunity knocks.

BY ROSS CLONEY

Tuesday the 14th, six in the morning and I woke up to find my 16-year-old self at the foot of my bed.

"I'm really sorry." She sniffed and then wiped her nose on her sleeve. "They said I really should do this as soon as I arrived."

"They?" I cautiously sat up in bed and glanced around. Aside from the impossible right in front of me, everything else seemed perfectly normal.

She looked off into the middle distance like she was dragging up a buried memory, "The … um … the Buddhist and the Nietzschian arguing at the end of the Universe."

I stared blankly. This must be a stress dream. A stalled research programme, an empty bank account and no idea of what to do with my future. No wonder I was dreaming about me at 16, looking just like I did when I first knew I wanted to be a physicist.

"There's a Buddhist at the end of the Universe?" The question was out of my mouth before I realized how insane it sounded.

My younger self sat on the edge of the bed. She was seeming more focused now.

"What you would recognize as a Buddhist. And a follower of what you would ascribe to Nietzsche. The last two viewpoints left after all the others fell away, eroded by time. They say this meeting is very important."

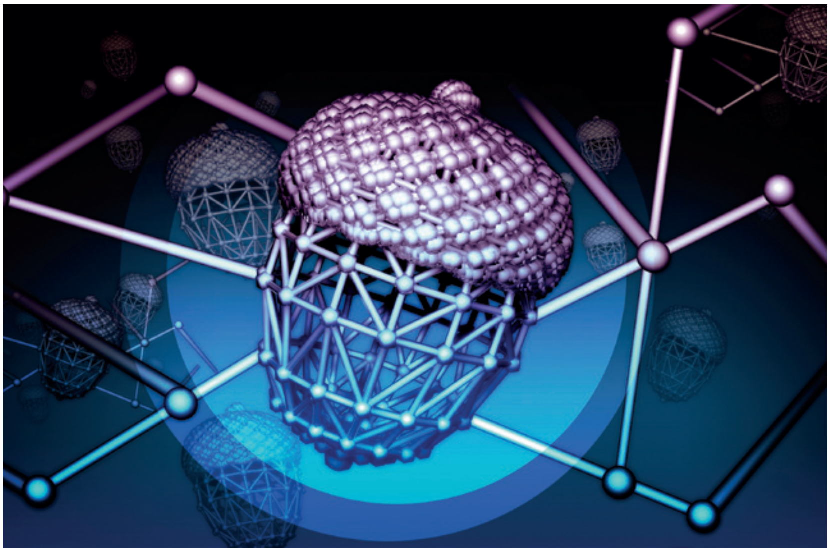
"So … there are only two people left?" Part of me wondered why I was interrogating a dream.

The focus in her eyes was intense now. "They are not singular entities. They … we are amalgamations of vast numbers of individuals, species, societies. Sentient worldviews, omega-level cultures at the very summit of technological progress."

My mouth soundlessly repeated what she said. This was straight out of the science fiction I used to read. Although that never mentioned Buddhists.

"And you want to talk to me?"

"We have as best we can stalled the end of existence but we cannot slow the heat death any longer. The dharma wheel wobbles on its axis. The will to power no longer has fresh ground to impress itself upon. Change is ceasing. And for all our aeons of arguing, we agree that this is undesirable. The forced cessation of samsara does not automatically earn nirvana. Nonexistence is the antithesis of life affirming."



JACEY

"And what does this have to do with me?" My suspicion and disbelief were steadily being overcome by curiosity. The bed felt firm beneath me, the air from the window had an autumnal chill to it. If this was a dream, it was the most vivid one I had ever had.

My younger self turned her head to look at my desk. She even had those stupid star earrings I wore until undergrad. Although hers were red, not the blue I remember having.

"You were writing a very interesting paper."

There it was, same place it had been for months, before the funding dried up and my motivation with it. 'Is a multiverse needed? New interpretations of old problems'. It wasn't even what the grant was on, just a thought game I had been playing around with.

"So something about my research is, or will be, key to stopping the heat death of the Universe?"

She smiled at me. "In the singular no. We are making three million four hundred and sixty-two thousand nine hundred and twelve interventions across a period of roughly 1 billion years and a span of 63.7% of what you would consider the observable Universe with varying degrees of direct or violent persuasion. To outside observers they would appear simultaneous."

"Oh." I suddenly felt very small and a little threatened. Then that phrase drifted back to me. *To outside observers.*

"Besides, the heat death cannot be avoided. This Universe will end. We seek a way to keep the wheel of dharma turning. To find a new will to power."

"I don't understand … you're not trying to alter the past to avoid your fate?"

"We are the summation of all that has come before but we are limited to our own timeline as much as you are. We cannot change the past. All that has happened has happened. But not all that has occurred has happened."

I looked at my younger self. *Not all that has occurred has happened.* Understanding was right at the edge of my awareness. Had I no memory of this meeting because it hadn't happened in the timeline of my past?

"You're trying to create … alternatives? You can't change the past but you can try to create divergent timelines, some of which will have a way for you to … escape."

"There is one timeline but much potential. We can stitch what could have been onto what is. For a while. But even a moment would be enough to start a new cycle."

My alternate stood up, smoothed out her dress and walked to the window. She was looking pale, almost immaterial. How long could a god-level-civilization-entity keep timelines stitched together?

"I still don't understand what my role in this is," I said.

"The same role an acorn plays in a forest," she said, smiling. Then a gust of wind and she was dust, slowly settling on the half-finished manuscript on my desk. ■

Ross Cloney *is a postdoctoral fellow based in Brighton, UK, working in genome stability. In his spare time he wants to clone dinosaurs for a completely original idea he's had for a theme park.*